THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *08/10/2013* **par :**

Leveraging social relevance: Using social networks to enhance literature access and microblog search

---

JURY

Claude CHRISMENT	Professeur, Université de Toulouse 3	Président du jury
Patrick GALLINARI	Professeur, Université Pierre et Marie Curie	Examinateur
Gabriella PASI	Professeur, Università di Milano Bicocca	Rapportrice
Eric GAUSSIER	Professeur, Université de Grenoble 1	Rapporteur
Lynda TAMINE	Professeur, Université de Toulouse 3	Directrice
Mohand BOUGHANEM	Professeur, Université de Toulouse 3	Co-directeur

---

École doctorale et spécialité :
MITT : Image, Information, Hypermedia
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)
Directeur(s) de Thèse :
Lynda TAMINE et *Mohand BOUGHANEM*
Rapporteurs :
Gabriella PASI et *Eric GAUSSIER*

Leveraging social relevance: Using social networks to enhance literature access and microblog search

Lamjed BEN JABEUR

October 7^{th}, 2013

Leveraging social relevance: Using social networks to enhance literature access and microblog search
Thesis submitted for the degree of Doctor of Philosophy
Thesis defended on October 8^{th}, 2013

Ph.D: Lamjed Ben Jabeur
Supervisor: Prof. Lynda Tamine, University of Toulouse 3 Paul Sabatier
Advisor: Prof. Mohand Boughanem, University of Toulouse 3 Paul Sabatier

v1.2 2013-12-06
Comments, corrections, and other feedback most welcome at:
{jabeur,tamine,boughanem}@irit.fr

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS,
Université Toulouse 3 Paul Sabatier,
118 route de Narbonne,
F-31062 Toulouse CEDEX 9

Dedicated to my family, I love you!

Acknowledgments

I would like to express my heartfelt gratitude to my supervisor, Professor *Lynda Tamine*, for the continuous academic and emotional support, her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I can't forget her hard times reviewing my thesis progress, giving me her valuable suggestions and made corrections. Her unflinching courage and conviction will always inspire me, and I hope to continue to work with her noble thoughts. My sincere gratitude goes to my advisor, Professor *Mohand Boughanem*, for the guidance, encouragement, suggestions and valuable corrections he has provided throughout my time as his student. His enthusiasm for research inspired me to continuously improve as a scientist.

I would also like to thank the examiners of my thesis, Professor *Gabriella Pasi*, University of Milano-Bicocca, and Professor *Eric Gaussier*, University of Joseph Fourier Grenoble 1, who provided encouraging and constructive feedback. I am grateful for their thoughtful and detailed comments. I'm equally thankful to Professor *Claude Chrisment*, University of Toulouse 3 Paul Sabatier, and Professor *Patrick Gallinari*, University of Pierre and Marie Curie Paris 6, for being part of the jury and for their brilliant comments and suggestions

This thesis was funded by *Quaero* project, and I would like to thank all contributors for their generous support. A heartfelt thanks to *Karen Pinel-Sauvagnat* and supportive colleagues at Generalized Information Systems team *SIG*, all the members of Toulouse Institute of Computer Science Research *IRIT*, Mathematics, Informatics & Telecommunications Toulouse Doctoral School *EDMITT* and University of Toulouse 3 Paul Sabatier for their instant help and kindness. Many thanks go to everyone who participated in this research study, *volunteers* who have participated in conducted user evaluation and administrators of *OSIRIM* platform for their assistance.

I would like thank my childhood friends *Ossama*, *Aymen* and *Nadhem*, for their support, encouragement and for those special moments that they have provided for me during vacances. A lot of thanks and gratitude to *Ossama*, *Abdessattar*, *Fatma*, *Ahmed*, aunt *Rachida* and uncle *Brahim* for their love and support. Special thanks go to *Zied* and *Faten* for their help during difficult times.

I would like to thank all my friends at IRIT lab including *Laure*, *Arlind*, *Imen*, *Hamdi*, *Cyril*, *Firas*, *Selma*, *Fatma*, *Manel*, *Akram*, *Dana*, *Madalina*, *Eya*, *Rafik*, *Bilel*, *Ismail*, *Feiza*, *Valentin*, *Hajer*, *Adel*, *Dana*, *Ihab*, *Anas*, *Aida*, *Bochra*, *Mariam*, *Mouna*, *Saad*, *Slim*, *Rahma* and *Nicolas*. All of you are great! I'm grateful for *Imed*, *Firas*, *Mohamed*, *Chokri*, *Youssef*, *Majid*, *Slah* and *Hela* as well as all my friends and teachers at *ISAMM*, *ISIMS* and *UPS* and everyone who helped me to success in my studies.

Lastly, and most importantly, I would like to thank my family for providing a loving environment for me. *Father*, I truly cannot thank you enough. I certainly would not be where I am today without your love and guidance since the day I was born. Thank you *Mother* (RIP) for sitting on my shoulder and guiding me through my life, instilling me with a strong passion for learning and a great motivation to go forwards to further success. Celebrate! Your kid has made it through the Ph.D! Thank you brother *Nabil*, I know how much you have sacrificed to help me. Wish you the best! Thank you *sisters*, *brothers*, *step-mother*, *sisters-in-law*, *nephews* and *nieces*. You believed in my dream and you helped me to realize it. Thank you infinitely, I love you so much!

Abstract

An information retrieval system aims at selecting relevant documents that meet user's information needs expressed with a textual query. For the years 1970-1980, various theoretical models have been proposed in this direction to represent, on the one hand, documents and queries and on the other hand to match information needs independently of the user. More recently, the arrival of *Web 2.0, known also as the* social Web, has questioned the effectiveness of these models since they ignore the environment in which the information is located. In fact, the user is no longer a simple consumer of information but also involved in its production. To accelerate the production of information and improve the quality of their work, users tend to exchange documents with their social neighborhood that shares the same interests. It is commonly preferred to obtain information from a direct contact rather than from an anonymous source. Thus, the user, under the influenced of his social environment, gives as much importance to the social prominence of the information as the textual similarity of documents at the query. In order to meet these new prospects, information retrieval is moving towards novel user centric approaches that take into account the social context within the retrieval process.

Thus, the new challenge of an information retrieval system is to model the relevance with regards to the social position and the influence of individuals in their community. The second challenge is produce an accurate ranking of relevance that reflects as closely as possible the importance and the social authority of information producers. It is in this specific context that fits our work. Our goal is to estimate the social relevance of documents by integrating the social characteristics of resources as well as relevance metrics as defined in classical information retrieval field.

We propose in this work to integrate the social information network in the retrieval process and exploit the social relations between social actors as a source of evidence to measure the relevance of a document in response to a query. Two social information retrieval models have been proposed in different application frameworks: literature access and microblog retrieval. The main contributions of each model are detailed in the following.

A social information model for flexible literature access We proposed a generic social information retrieval model for literature access. This model represents scientific papers within a social network and evaluates their importance according to the position of respective authors in the network. Compared to previous approaches, this model incorporates new social entities represented by annotators and social annotations (tags). In addition to co-authorships, this model includes two other types of social relationships: citation and social annotation. Finally, we propose to weight these relationships according to the position of authors in the social network and their mutual collaborations.

A social model for information retrieval for microblog search We proposed a microblog retrieval model that evaluates the quality of tweets in two contexts: the social context and temporal context. The quality of a tweet is estimated by the social importance of the corresponding blogger. In particular, blogger's importance is calculated by the applying PageRank algorithm on the network of social influence. With the same aim, the quality of a tweet is evaluated according to its date of publication. Tweets submitted in periods of activity of query terms are then characterized by a greater importance. Finally, we propose to integrate the social importance of blogger and the temporal magnitude tweets as well as other relevance factors using a Bayesian network model.

Résumé

L'objectif principal d'un système de recherche d'information est de sélectionner les documents pertinents qui répondent au besoin en information exprimé par l'utilisateur à travers une requête. Depuis les années 1970-1980, divers modèles théoriques ont été proposés dans ce sens pour représenter les documents et les requêtes d'une part et les apparier d'autre part, indépendamment de tout utilisateur.

Plus récemment, l'arrivée du *Web 2.0* ou le *Web social* a remis en cause l'efficacité de ces modèles du fait qu'ils ignorent l'environnement dans lequel l'information se situe. En effet, l'utilisateur n'est plus un simple consommateur de l'information mais il participe également à sa production. Pour accélérer la production de l'information et améliorer la qualité de son travail, l'utilisateur échange de l'information avec son voisinage social dont il partage les mêmes centres d'intérêt. Il préfère généralement obtenir l'information d'un contact direct plutôt qu'à partir d'une source anonyme. Ainsi, l'utilisateur, influencé par son environnement socio-cultuel, donne autant d'importance à la proximité sociale de la ressource d'information autant qu'à la similarité des documents à sa requête. Dans le but de répondre à ces nouvelles attentes, la recherche d'information s'oriente vers l'implication de l'utilisateur et de sa composante sociale dans le processus de la recherche.

Ainsi, le nouvel enjeu de la recherche d'information est de modéliser la pertinence compte tenu de la position sociale et de l'influence de sa communauté. Le second enjeu est d'apprendre à produire un ordre de pertinence qui traduise le mieux possible l'importance et l'autorité sociale. C'est dans ce cadre précis, que s'inscrit notre travail. Notre objectif est d'estimer une pertinence sociale en intégrant d'une part les caractéristiques sociales des ressources et d'autre part les mesures de pertinence basées sur les principes de la recherche d'information classique.

Nous proposons dans cette thèse d'intégrer le réseau social d'information dans le processus de recherche d'information afin d'utiliser les relations sociales entre les acteurs sociaux comme une source d'évidence pour mesurer la pertinence d'un document en réponse à une requête. Deux modèles de recherche d'information sociale ont été proposés à des cadres applicatifs différents : la

recherche d'information bibliographique et la recherche d'information dans les microblogs. Les importantes contributions de chaque modèle sont détaillées dans la suite.

Un modèle social pour la recherche d'information bibliographique. Nous avons proposé un modèle générique de la recherche d'information sociale, déployé particulièrement pour l'accès aux ressources bibliographiques. Ce modèle représente les publications scientifiques au sein d'réseau social et évalue leur importance selon la position des auteurs dans le réseau. Comparativement aux approches précédentes, ce modèle intègre des nouvelles entités sociales représentées par les annotateurs et les annotations sociales. En plus des liens de coauteur, ce modèle exploite deux autres types de relations sociales : la citation et l'annotation sociale. Enfin, nous proposons de pondérer ces relations en tenant compte de la position des auteurs dans le réseau social et de leurs mutuelles collaborations.

Un modèle social pour la recherche d'information dans les microblogs. Nous avons proposé un modèle pour la recherche de tweets qui évalue la qualité des tweets selon deux contextes: le contexte social et le contexte temporel. Considérant cela, la qualité d'un tweet est estimé par l'importance sociale du blogueur correspondant. L'importance du blogueur est calculée par l'application de l'algorithme PageRank sur le réseau d'influence sociale. Dans ce même objectif, la qualité d'un tweet est évaluée selon sa date de publication. Les tweets soumis dans les périodes d'activité d'un terme de la requête sont alors caractérisés par une plus grande importance. Enfin, nous proposons d'intégrer l'importance sociale du blogueur et la magnitude temporelle avec les autres facteurs de pertinence en utilisant un modèle Bayésien.

Publications

International journals articles

1. Laure Soulier, Lamjed Ben Jabeur, Lynda Tamine, Wahiba Bahsoun. On Ranking Relevant Entities in Heterogeneous Networks Using a Language-Based Model In *Journal of the American Society for Information Science and Technology (JASIST)*, Wiley, Vol. 64 N. 3, p. 500-515, mars 2013.

Books parts

1. Lynda Tamine, Lamjed Ben Jabeur, and Wahiba Bahsoun. On using social context to model information retrieval and collaboration in scientific research community . In *Community-Built Database: Research and Development*, chapter 6, pages 133–155. Springer, mai 2011.

International conferences articles

1. Lamjed Ben Jabeur, Firas Damak, Lynda Tamine, Guillaume Cabanac, Karen Pinel-Sauvagnat, and Mohand Boughanem. IRIT at TREC Microblog 2012: Adhoc Task. In*Text REtrieval Conference (TREC)*, Gaithersburg, USA, novembre 2012. National Institute of Standards and Technology (NIST).

2. Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Featured tweet search: Modeling time and social influence for microblog retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI 2012)*, 4-7 December 2012 , Macau.

3. Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Active microbloggers: Identifying influencers, leaders and discussers in microblogging networks. In *International Symposium on String Processing and Information Retrieval (SPIRE 2012)*, 21-25 October , Cartagena de Indias, Colombia, 2012.

4. Laure Soulier, Lamjed Ben Jabeur, Lynda Tamine, and Wahiba Bahsoun. BibRank: a Language-Based Model for Co-Ranking Entities in Bibliographic

Networks. In *Joint Conference on Digital Libraries (JCDL 2012)*, June 10-14, Washington, DC, 2012.

5. Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Uprising microblogs: A Bayesian network retrieval model for tweet search. In *ACM Symposium on Applied Computing (SAC)*, Riva del Garda (Trento), Italy, mars 2012.

6. Firas Damak, Lamjed Ben Jabeur, Guillaume Cabanac, Karen Pinel-Sauvagnat, Lynda Tamine, and Mohand Boughanem. IRIT at TREC Microblog 2011. In *Text REtrieval Conference (TREC)*, Gaithersburg, USA, novembre 2011. National Institute of Standards and Technology (NIST).

7. Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. A social model for Literature Access: Towards a weighted social network of authors. In *International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO)*, Paris, France, avril 2010. Centre de hautes études internationales d'Informatique Documentaire (C.I.D.).

8. Lynda Tamine, Lamjed Ben Jabeur, and Wahiba Bahsoun. An exploratory study on using social information networks for flexible literature access. In *Flexible Query Answering (FQAS)*, Roskilde, volume 5822 of Lecture Notes in Computer Science, pages 88–98, octobre 2009. Springer.

National conferences articles

1. Laure Soulier, Lamjed Ben Jabeur, Lynda Tamine, and Wahiba Bahsoun. Modèle de langue pour l'ordonnancement conjoint d'entités pertinentes dans un réseau d'informations hétérogènes. In *NFormatique des Organisations et Systemes d'Information et de Decision (INFORSID 2012)*, 29-31 Mai, Montpellier, 2012.

2. Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Intégration des facteurs temps et autorité sociale dans un modèle bayésien de recherche de tweets. In *Conférence francophone en Recherche d'Information et Applications (CORIA)*, Bordeaux, 2012.

3. Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter. In *Conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI)*, Grenoble, octobre 2011. IMAG.

4. Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem. Un modèle de Recherche d'Information Sociale pour l'Accès aux Ressources Bibliographiques : Vers un réseau social pondéré. In *Atelier REcherche et REcommandation d'information dans les RESeaux sOciaux à INFORSID 2010*, Marseille, pages 37–49, mai 2010. Association INFORSID.

5. Lamjed Ben Jabeur and Lynda Tamine. Vers un modèle de Recherche d'Information Sociale pour l'accès aux ressources bibliographiques. In *Conférence francophone en Recherche d'Information et Applications (CORIA)*, Sousse, Tunisie, pages 325–336, mars 2010. Centre de Publications Universitaires.

Contents

Chapter 1

Introduction

1.1 Emergence of the Social Web

The rise of social Web has changed the landscape of the World Wide Web. This concept was introduced in 1996 by Rheingold. Besides the principle aim of the Web to ensure open access to information, the new generation of social websites has enabled users to communicate effectively with each others (Rheingold, 2000). The first social networking services such as Classmates[1] (1995) and SixDegrees[2] (1996) transformed the structure of the Web from a hypertext environment that links data to a *"Web of people"* environment that connects family, friends and colleagues.

With the launch of the first blogging service OpenDiary[3] in 1998, Internet users were given the opportunity to publish their own content on the Web. They can interact with each other and post comments on published content. Interaction between users is promoted later by Wiki platforms, namely Wikipedia[4] (2001). Such service enabled online communities, on the first hand, to exchange their knowledge, and on the other, to efficiently collaborate online. The popularity of these websites is followed by the growth of other social networking services such as Myspace[5] (2003), Facebook[6] (2004), LinkedIn[7] (2006) and Twitter[8] (2006). These websites have not only instated a novel practice on the Web but also introduced a new life style for Internet generation. Social networking services

[1] http://www.classmates.com/
[2] http://www.sixdegrees.com/
[3] http://www.opendiary.com/
[4] http://www.wikipedia.org/
[5] http://www.myspace.com/
[6] http://www.facebook.com/
[7] http://www.linkedin.com/
[8] http://www.twitter.com/

have widely impacted communication, education, and entertainment as well as commercial, financial and governmental services.

With the exponential growth of the social Web, the role of Internet users has been transformed from passive information consumers to active producers. Over and above professional content edited by webmasters, Web users made together a substantial effort to produce and publish their own content (Vickery and Wunsch-Vincent, 2007). They contribute in different ways and with various formats to enrich the Web experience. Data published by users is known as User Generated Content (UGC). It covers (*i*) original or compiled materials that users make available over blogs, wikis and media sharing services (*e.g.*, YouTube[9], Flikr[10], *etc*); (*ii*) feedback and metadata such as comments, review, rates, and tags; (*iii*) and finally social network data including public profiles, social network structure and user interactions.

Recent statics of social networking services[11] shows high user participation rate as well an incredible amount of UGC published daily. According to eMarketer, the number of active social network users, presented in table 1.1, is estimated to 1.43 billion in 2012. This number will reach 1.85 billion in 2014. The same source reports that 68% of internet users in United States are costumers of UGC as presented in table 1.2. Blog remains the most popular social networking service attracting 76% of Internet users. Social networking and videos sharing websites attract, respectively, 50.5% and 47.2% of Internet users.

	2011	2012	2013	2014
Social network users	1.2	1.43	1.66	1.85
% change	23.1%	19.2%	16.6%	11.6%

Table 1.1: Social network users worldwide (billions users). Source: eMarketer, February, 2012.

Table 1.3 lists most popular social networking websites[12] and corresponding number of registered users and monthly active users. Facebook claims to be the largest virtual community with 1 billion active users according to statistics published in September 2012[13]. About 130 million status updates, 300 million photos and 730 comments are daily published on Facebook[14] (2011). Twitter, the most popular microblogging service, claims 200 million active users on Mars 2013 and about 400 posts a day[15]. Other social networking services show also an increasing number of daily published UGC including reviews, comments and tagged objects.

[9] http://www.youtube.com
[10] http://www.flickr.com/
[11] https://www.emarketer.com/coverage/socialmedia
[12] http://en.wikipedia.org/wiki/List_of_virtual_communities_with_more_than_100_million_active_users
[13] http://newsroom.fb.com/News/One-Billion-People-on-Facebook-1c9.aspx
[14] http://www.onlineschools.org/visual-academy/facebook-obsession/
[15] https://blog.twitter.com/2013/celebrating-twitter7

	2008	2009	2010	2011	2012	2013
User-generated video	36.0%	39.8%	42.5%	44.8%	47.2%	49.2%
Social networking	41.2%	44.2%	46.9%	49.1%	50.5%	51.8%
Blogs	54.0%	58.0%	61.0%	64.0%	67.0%	69.0%
Wikis	33.9%	36.6%	39.0%	41.0%	42.6%	43.9%
UGC consumers	60.0%	62.0%	64.0%	66.0%	68.0%	70.0%

Table 1.2: US User Generated Content consumers. Source: eMarketer, January, 2009.

Name	Registered	Monthly active	Date
Facebook	1000+	1000	Oct. 2012
Skype	663+	280	Jan. 2013
Google+	500+	235	Dec. 2012
Twitter	500+	200	Dec. 2012
LinkedIn	225+	160	Jan. 2012
Windows Live	100+	100	Dec. 2012

Table 1.3: Popular social networking services (million users)

1.2 Towards Social Information Retrieval

The emergence of UGC on Internet has resulted in a massive source of information that grows steadily in quantity and quality. On the other hand, the number of search queries has considerably increased. Google reports that the number of search queries has grumped from 9.8 thousand queries per day in 2011 to 3 billion searches a day in August 2012[16]. Twitter is handling over 1.6 billion queries every day by April 2011 [17]. Facebook processes more than 1 billion daily search queries in September 2012 [18]. Certainly, the change of user role from information consumer to content producer has impacted his need of information and led to this massive amount of search queries. In fact, users become greedier for fresh and accurate information than before. Being aware of UGC availability and the original content that may provide, users express a desire to access to this type of data. In reality, what would make such content even more valuable is that it represents both collective knowledge and user interactions.

UGC provides exhaustive information that may have not been included yet in Web pages maintained by professionals. For instance, phone constructor website may not be useful to check if a new application is supported by some Smartphone. A trusted blog review from technology expert or a reply in a

[16]http://www.google.com/competition/howgooglesearchworks.html
[17]http://engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html
[18]http://www.theverge.com/2012/9/11/3317720/facebook-billion-search-queries-a-day

questioning-answer website would be practical in this case. Furthermore, a user would be particularly interested in some Smartphone if a similar friend has already purchased it. UGC and user relationships could be exploited to enhance information access and provide quality data that satisfies user's needs of information and helps him to accomplish his task.

The motivation behind the use of UGC within information access and retrieval system, typically for Web search, is to take advantage of the *"Wisdom of Crowds"* and leverage the search accuracy. The *"Wisdom of Crowds"* concept, introduced by Surowiecki (2005), refer to the collective intelligence elaborated by Internet users who collaborate to tag, rate and review Web resources via Wikis and blogs. As illustrated in figure 1.1, these interactions are useful for accessing to web resources. It allows retrieval systems to gather user feedback and thus present accurate results to his information need.

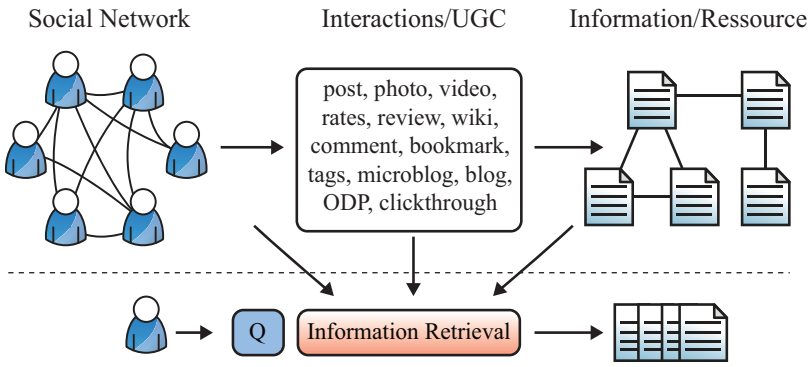


Figure 1.1: Using UGC to enhance infromation retrieval

Social networks and UGC could be integrated along retrieval processes as information source for relevance feedback and personalized access. For instance, user queries may be expanded using Wikis (Koolen et al., 2009), ODP collaborative directory (Bai et al., 2007) and tagging information (Heymann et al., 2008b). New published Web pages could be instantly detected thanks to blogs and microblogging stream (Rowlands et al., 2010; Dong et al., 2010). Click through data such as query logs, clicks, bookmarks may be used for ranking Web resources (Joachims, 2002; Xue et al., 2004). The main difference between these approaches and traditional information retrieval models focusing on information entities (*i.e.*, documents and terms) is to take into account the social context of Web resources.

Despite the promising role of UGC in information search and access, retrieval processes should focus on user as a primer unit of information. In fact, Internet users search for relevant information with regards to their information needs but, at the same time, wonder which person has published this information? Is he a reliable source of knowledge? How other people do think about? These

questions show that users care about the quality of information as much as the quality of persons behind it, typically their endorsement in the social network. Accordingly, the concept of relevance within the social Web is extended to cover documents as well as actors in interaction at both information and producing consuming levels. In other words, information relevance is defined by the importance of related people in the social network and vice versa. Within this view, the use of theoretical foundations of social networks in information retrieval and access becomes necessary to achieve new information needs where people are as much worthy as the information itself.

Social Information Retrieval (SIR) is proposed in this context as a novel research area that bridges information retrieval and social networks analysis research areas in order to enhance retrieval processes by means of social usage of information (Korfiatis et al., 2006; Kirchhoff et al., 2008). In particular, the social context is inferred from the analysis of unstructured communication between users (Kleinberg, 2008) as well as user profiles and interactions. Indeed, analyzing what people say, share and annotate allows identifying what would better meet their information needs. In practice, social information retrieval systems consider the social network as well as implicit and explicit indicators of information interest such as tagging, rating and friend activities in order to estimate the relevance of information items (Amer-Yahia et al., 2007). Social network analysis methods are applied at this aim to identify important persons in the social network then evaluate the relevance of information items respectively to the position of related people in the social network.

1.3 Challenges for Social Information Retrieval

Social information retrieval has captured in the latest years the interest of scientific research and industrials. A significant effort is invested by information retrieval community to design, implement and evaluate new generations of information retrieval systems where the social Web is a central object of study. There are remaining challenges for social information retrieval to overcome. We discuss in what follows the main challenges faced to:

Volume and sparsity. As discussed above, the emergence of the social Web has led to a huge quantity of user generated data. Obviously, data availability may improve the effectiveness of information retrieval systems. However, this has a paradoxical outcome. In fact, retrieval systems should be able to process this amount of data and make it usable. The challenge covers technological aspect of information processing such as indexing and searching as well as conceptual and methodological aspects. The first issue addresses storage, access and large-scale analysis of massive quantities of information, or, as commonly called, *"Big data"* (Zikopoulos et al., 2011; Dean and Ghemawat, 2008). The second issue addresses rather the question of what knowledge can be learned from social networks

and user generated data (Du et al., 2007; Mislove et al., 2007). Furthermore, data availability is conversely accompanied to sparsity typically over the social network structure. This may restrict the effectiveness of information retrieval systems on large-scale social networks where user interactions are irregular or some information is missed by privacy concern.

Compositional and structural divergence. Each social networking service proposes an original network structure that differentiates it to competitors. For instance, *friendship* associations equally connect friends in Facebook. Twitter proposes one-way relationships known as *followership*. Google+ adopts however another approach where social connections are classified into *confidence circles* (*e.g.*, family, colleagues, friends and acquaintances). In addition, the social network may involve different types of entities according to the networking activities. In Wiki social networks, two types of entities are involved: authors and articles. Social bookmarking networks involve more entities including users, documents and tags. This diversity of social network structures brings more challenges to social information retrieval. Approaches proposed in this context may support the structural divergence of social networks.

Evaluation of the social context. The evaluation of the social context helps to identify central entities in the social network. The definition of the social relevance depends however on the social purpose of the networking application as well as the social network structure. For example, important actors in Wikis are defined by experts with valuable contributions on some topic and who received at the same time less criticisms. In the case of media sharing networks, the social relevance is assimilated to the popularity of the user. Beside these two properties, the social relevance may be defined by the authority, the trust and the influence of persons on the social network. An information retrieval system must identify what property would better reflect the social relevance networking application and proposes convenient social network structure and metrics that enables to evaluate it.

Composite definition of relevance. Actors interactions in the social networks determine the relevance of connected resources. Beside this, other factors with respect of the networking activity may contribute to resources relevance. For example, the location and the proximity of an object would determine its relevance in a geographic tagging application as the main networking activity require. Timeline and freshness would be important indicators in a collaborative news headline system. Other criterion such as video quality in media sharing service would determine the social relevance of content. The challenge for social information systems is to define, model and integrate these factors in order to compute a global relevance of resource.

1.4 Research questions and focus

This thesis focuses on the problem of relevance definition within social network, specifically the evaluation of social importance of network actors. Two main research questions are being addressed:

1. What social network structure does define the social importance of actors?
   (a) What entities, actors and interactions do represent the social context?
   (b) Which network properties do reflect the social importance of actors?
   (c) How to evaluate the social importance of actors in social networks?
2. How to integrate the social context into the retrieval process?
   (a) What factors do contribute to information relevance?
   (b) How to combine relevance factors into an integrated retrieval process?

1.5 Contributions

The main contributions of this thesis consist on integrating the social network properties in the information retrieval process. In particular, we exploit the social relations between actors as a source of evidence to evaluate the relevance of documents. Two information scenarios are addressed in this context. In the first scenario, the retrieval process is conducted over traditional documents (*i.e.*, scientific articles). Relevance in this case is estimated based on the social network of respective authors and annotators. The second scenario target however social networking environments where UGC (*i.e.*, microblogs) is the focus of the retrieval task. Accordingly, we propose two social information retrieval models addressing two different application frameworks: (*i*) a social information retrieval model for literature access and (*ii*) a social information retrieval model for microblogs. The two contributions are summarized in the following.

Social information retrieval model for literature access. We propose a generic model for social information retrieval deployed particularly for literature access. This model represents scientific publications with social networks and evaluates their importance according to the position of respective authors in the social network. Compared to previous approaches, this new model incorporates new social entities such as annotators and tags. In addition to co-authorship, this model integrates other types of social relationships such as citations and social annotations. Finally, we propose to evaluate these relationships from the position of related actors in the social network and their mutual interactions.

Social information retrieval model for microblogs. We propose a social model for tweet search that, first, identifies important microbloggers in the social networks then evaluates the relevance of tweets in respect of the position of related microbloggers in the social network as well as microblogging features and the

temporal relevance of tweet. In particular, influencers, leaders and discussers are investigated as key microbloggers in the social network. For this aim, we introduce a social network model for microbloggers that represents microbloggers using multigraphs and integrates different types of associations including followerships, retweets and mentions. Three link analysis algorithms are proposed based on this social network model in order to identify network influencers, leaders and discussers.

Furthermore, we propose to evaluate tweet relevance according to its date of publication. The distribution of terms are thus analyzed in order identify activity periods of a query topic. Tweet submitted in accordance with a query event are presumed relevant.

Finally, we propose to integrate the topical relevance, the social importance of a microblogger and the temporal relevance of the tweets into an integrated Bayesian framework. Two topologies of Bayesian network models for tweet search are proposed in this context based on a Bayesian inference network and a belief network.

1.6 Thesis overview

This thesis is structured into 8 chapters. The content of each chapter is described in what follows:

Chapter 1 gives an overview of this thesis. Research questions and main continuations are also presented in this section.

Chapter 2 presents a short introduction to basic concepts of information retrieval, social network analysis and social information retrieval.

Chapter 3 focuses on literature access and presents an overview of related work that applies social network approaches in this domain. First, different models of scientific social network structures are discussed. Thereafter, we discuss social-based information retrieval models for literature access and retrieval. Finally, we propose a comparison of social services proposed within main digital libraries and academic search engines.

Chapter 4 review related work on information retrieval over microblogs. First, microblog properties and the social network characteristics are presented. Second, we discuss the main information retrieval tasks over microblogs. After that we focus on social retrieval approaches for microblog search. Finally, we discuss the evaluation of microblogging retrieval systems.

Chapter 5 introduces our social model for literature access. First, he qualitative and the quantitative model of the social network are presented. Afterward, we detail the computation of social importance scores then the estimation of the global relevance of scientific publications.

Chapter 6 introduces a social network model for identifying key actors in microblogging networks. First, a formal model for microblog social network is presented. Subsequently, three link analysis algorithms are proposed in order to identify influencers, leaders and discussers.

Chapter 7 presents two integrated models for microblog search based on inference Bayesian networks and belief Bayesian networks. For each model, we first present the topology of Bayesian network model then we focus on the query evaluation process. The two proposed models are finally compared with state-of-the-art approaches.

Chapter 8 concludes this thesis, discusses findings and outlines future work.

Chapter 2

Background

We present in this chapter basic concepts and definitions that will be used throughout this thesis. First, we will introduce the fields of information retrieval and social network analysis. Afterward, we present the social information retrieval research area which resulted from the combination of these two research fields.

2.1 Information retrieval

Information Retrieval (IR), one of early research domain in computer science, has proposed first automatic solutions for text storage and search (Luhn, 1957; Bush, 1979). Salton and McGill (1983) defined information retrieval system as the set of processes that provide user with information.

> *An information retrieval system is an information system, that is, a system used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations.*

The main purpose of an information retrieval system is to stratify user's information need. This need, usually formulated using a textual query, is motivated by a real world task. For instance, a student in biology preparing a report about genetics may think of *"Genetics, DNA, and Heredity"* as keywords to express his information need. Nevertheless, information retrieval deals with information and ideas instead of words and phrases. One challenge ahead is to understand the information need of the users behind these few words and provides useful items of information, qualified as *"relevant"*, that help user to accomplish his task.

We describe in this section basic concepts of information retrieval systems then we present an overview of information retrieval models and evaluation measures.

2.1.1 Basic concepts

An information retrieval system is comprised of 3 main processes: indexing, retrieval and ranking. These processes are more or less complex depending on the retrieval task. Figure 2.1 illustrates a typical architecture of an information retrieval system.

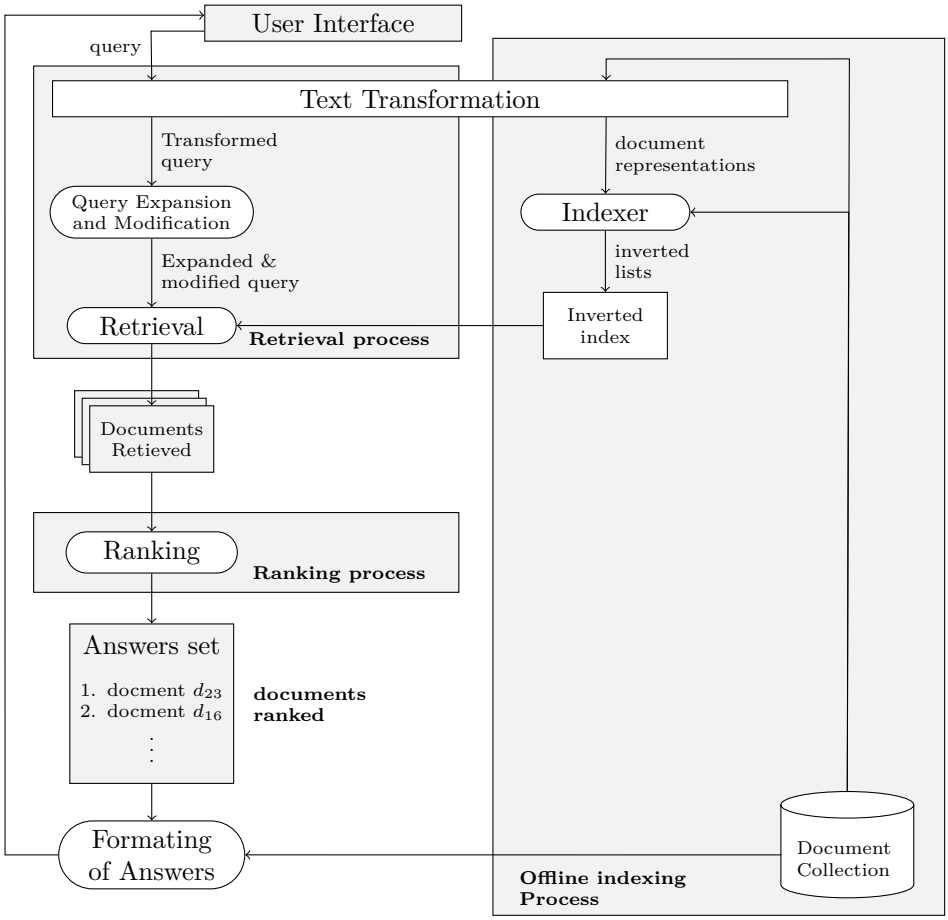


Figure 2.1: Information retrieval system (Baeza-Yates and Ribeiro-Neto, 2011)

Indexing process is performed once offline at the beginning of an information retrieval cycle. During this process, several *text transformation* and normalization methods applied to documents. First, the text of the document is split into tokens, which is equivalent to words. Useless words, known as *stopwords*, are then removed. *lemmatization* and *stemming* (Porter, 1997; Pirkola, 2001) are used to transform words with similar meaning into a common base form.

For instance, stemming process transforms *"waiting"* and *"waited"* to the root word *"wait"*. Next, documents and terms are represented using a common data structure called *"inverted index"* (Salton et al., 1983; Knuth, 1998; Zobel et al., 1998). This structure ensures a fast access to document collection by mapping terms to the set of documents where they appear.

Retrieval process aims to select relevant documents that cover user's information needs. This process depends on document representation, user's information needs and user preferences (*e.g.*, language, date, format, *etc*). Queries are in fact text-based representations for user's information needs (Belkin and Croft, 1992). Thereby, textual transformation previously applied on documents should be applied to the query too. However, query may be expanded or modified to support user preferences and relevance feedback (Harman, 1988; Robertson, 1991). At the end of the retrieval process, a list of retrieved documents that contain at least one term of the query, either in original or expanded form, is compiled. This list includes candidate relevant documents with respect to the query.

Ranking process assigns a relevance score to documents in the retrieved set respectively to their similarity to the query. An *answer set* is compiled where documents are ranked by decreased relevance score or by another criterion that the user may select. The answer set is finally formatted by adding document title and abstract before being restituted in the user interface.

2.1.2 Information retrieval models

An information retrieval model is a theoretical support that represents documents and queries, and defines a ranking strategy for retrieved documents. An information retrieval model is modeled with a quadruple $[\mathcal{D}, \mathcal{Q}, \mathcal{F}, \mathcal{R}(q_i, d_j)]$ (Baeza-Yates and Ribeiro-Neto, 2011) where : $\mathcal{D}$ is a set of logical views of documents; $\mathcal{Q}$ is a set of logical views of user information needs, called queries; $\mathcal{F}$ is a modeling framework for documents and queries; $\mathcal{R}(q_i, d_j)$ is a ranking function with $q_i \in \mathcal{Q}$ and $d_j \in \mathcal{D}$. Proposed information retrieval models in the literature address three main characteristics of documents including text, links and multimedia. Figure 2.2 presents a taxonomy for information retrieval models based on these properties.

In the category of text-based information retrieval models, three families of models are identified with respect to an implemented mathematical framework. First, Boolean models represent documents with a set of terms. Set theory operations are extended to perform document retrieval (Lancaster and Fayen, 1973; Salton et al., 1983; Fox and Sharan, 1986). Second, Vector Space models represent documents in a multidimensional space where each term corresponds to one dimension in the space (Salton et al., 1975; Deerwester, 1988; Kwok,

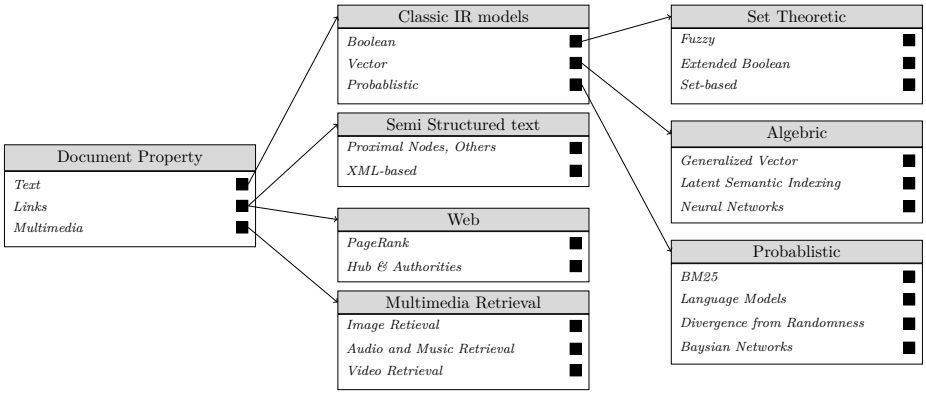


Figure 2.2: A taxonomy of Information Retrieval models (Baeza-Yates and Ribeiro-Neto, 2011)

1989). Finally, probabilistic models define probability distribution of each term and assimilate relevance as the document-query likelihood (Robertson et al., 1995; Ponte and Croft, 1998; Robertson and Walker, 1994; Ribeiro and Muntz, 1996).

Besides classic information retrieval models that mainly interest in textual properties, semi-structured text retrieval models has investigated the structure of documents (Perlman, 1993; Kotsakis, 2002). These models ensure different retrieval granularity and support complex query constraints on both content and structure. From another point of view, Web retrieval models exploit links as a key feature for identifying quality resources on the Web. PageRank (Brin and Page, 1998)and HITS (Kleinberg, 1999) are the two main models proposed in this category.

Finally, multimedia retrieval models propose a new alternative for retrieving non-textual data such as image, audio and video (Bird et al., 1996; Datta et al., 2008; Rueger, 2010). As these objects were originally expressed by a set of bits, a special index is built using textual and numeric properties such as media description, colors, shapes and textures.

We focus in what follows on main information retrieval model that we exploit in this thesis namely the Bayesian network model and PageRank model

2.1.2.1 Bayesian network model

Bayesian networks are graphical formalisms that model random variables and causal relationships between them (Pearl, 1985, 1988; Jensen, 2001). In particular, a Bayesian network is acyclic directed graph $G = (X, E)$ where the set of nodes X represents random variables and the set of edges $E = X \times X$ represents

conditional dependencies between them. Figure 2.3 illustrates an example of a Bayesian network graph.

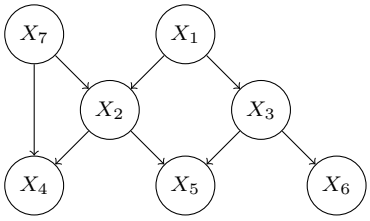
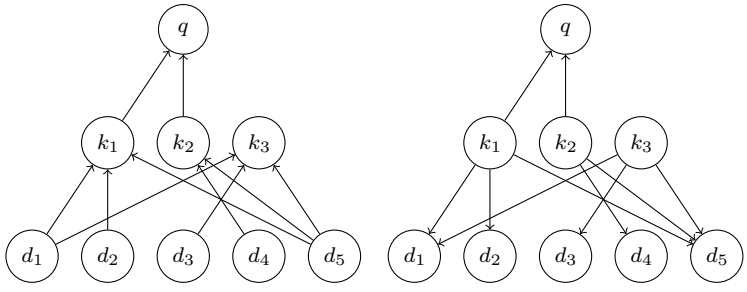


Figure 2.3: An example of a Bayesian network

Let $pa(X_i)$ be the set of parent (predecessor) nodes of a random variable X_i. The joint probability $P(X)$ for all variables X_i is computed as:

$$P(X) = P(X_1, X_2, \ldots, X_n) = \prod_{\forall X_i} P(X_i|pa(X_i)) \tag{2.1}$$

Based on this formalism, two basic Bayesian network models are proposed for information retrieval. First, inference networks model interprets probability from a statistical point of view (Turtle and Croft, 1990, 1991). The topology of this network model is illustrated in figure 2.4(a). Second, belief networks interpret probability as a degree of belief independently from statistical experiments (Ribeiro and Muntz, 1996; Silva et al., 2000; de Cristo et al., 2003). The topology of this model is illustrated in figure 2.4(b).



(a) Inference network model (b) Belief network model

Figure 2.4: Toplogies of inference and belief networks

Both inference network model and belief network models represent terms, documents and queries as random variables:

– Each term in the index is modeled by a random variable $k_i \in \{0, 1\}$. The event $k_i = 1$, simplified with k_i, denotes that term k_i is observed. The complementary event $k_i = 0$, simplified with $\bar{k}_i$, denotes that term k_i is

not observed. With k terms present in the index, it exists 2^k possible term configurations representing documents and queries. A term configuration is represented by a vector of random variables $\vec{k} = (k_1, k_2, \ldots, k_n)$ where each variable indicates if the corresponding term is observed. For example, an index of 2 terms k_1 and k_2 presents $2^2 = 4$ configurations represented by the following set $\mathscr{C} = \{(k_1, k_2), (k_1, \bar{k_2}), (\bar{k_1}, k_2), (\bar{k_1}, \bar{k_2})\}$. The event of observing a particular configuration $\vec{k} = (k_1, k_2, \ldots, k_n)$ is noted $\vec{k}$.

- Each document is modeled by a random variable $d_j \in \{0, 1\}$ with two possible values 0 or 1. The event $d_j = 1$, simplified with d_j, denotes that the document d_j is observed. Obviously, observing a document in a retrieval process means that this document is relevant to the query. On the other hand, the event $d_j = 0$, simplified with $\bar{d_j}$, denotes that document d_j is not observed.
- Each query is represented by a random variable $q \in \{0, 1\}$. The two events of observing the query ($q = 1$) or not observing the query ($q = 0$) are noted q and $\bar{q}$, respectively.

We detail in what follows the query evaluation process for each type of Bayesian network.

Inference network model. As shown in figure 2.4(a), inference network represents a query as a root node (Turtle and Croft, 1990, 1991). This node points to respective terms in the query. Each term points to the document where it appears. Accordingly, inference network topology expresses the probability of observing query terms in a document. Document relevance with respect to the query is assimilated to the probability $P(d_j \wedge q)$ of observing both document d_j and query q. This probability is developed by applying Bayes' rules as follows:

$$P(d_j \wedge q) = \sum_{\forall \vec{k}} P(q \wedge d_j | \vec{k}) P(\vec{k}) \tag{2.2}$$

$$= \sum_{\forall \vec{k}} P(q \wedge d_j \wedge \vec{k}) \tag{2.3}$$

$$= \sum_{\forall \vec{k}} P(q | d_j \wedge \vec{k}) P(d_j \wedge \vec{k}) \tag{2.4}$$

$$= \sum_{\forall \vec{k}} P(q | \vec{k}) P(\vec{k} | d_j) P(d_j) \tag{2.5}$$

$$P(\overline{d_j \wedge q}) = 1 - P(d_j \wedge q) \tag{2.6}$$

Notice that $P(q|d_j \wedge \vec{k})$ in equation 2.4 is transformed to $P(q|\vec{k})$ because q and d_j are d-separated given $\vec{k}$. $(\overline{d_j \wedge q})$ is the complement of $(d_j \wedge q)$. The sum of the two probabilities is therefore equal to 1, which argument equation 2.6.

Once index terms k_i are d-separated given d_j, the probability $P(\vec{k}|d_j)$ could be computed as the product of the probability of observing or not observing each

term k_i haven document d_j. Thus, equation 2.5 is rewritten as follows.

$$P(d_j \wedge q) = \sum_{\forall \vec{k}} P(q|\vec{k})P(d_j) \left(\prod_{\forall k_i | on(k_i,\vec{k})=1} P(k_i|d_j) \prod_{\forall k_i | on(k_i,\vec{k})=0} P(\bar{k_i}|d_j) \right) \tag{2.7}$$

$$P(\overline{d_j \wedge q}) = 1 - P(d_j \wedge q) \tag{2.8}$$

Where $on(k_i, \vec{k}) = 1$ if $k_i = 0$ according to $\vec{k}$, otherwise $on(k_i, \vec{k}) = 0$.

Belief network model. Belief network model (Ribeiro and Muntz, 1996; Silva et al., 2000; de Cristo et al., 2003) presents a similar topology than the inference network model with a slight difference on edge directions as shown in figure 2.4(b). The Belief network model considers term nodes as network roots conversely to the inference network model where query node is the only graph node. Terms point to query node and document nodes. This topology supports symmetric representations of documents and queries. Relevance probability is assimilated to the degree of overlap provided by the two concepts: document and query. The belief network model assimilates the relevance of document d_j with respect to query q with the probability of observing the document given the query $P(d_j|q)$. This probability is given by:

$$P(d_j|q) = \frac{P(d_j|q)}{P(q)} \tag{2.9}$$

Having $P(q)$ is constant for all documents, $P(d_j|q)$ is proportional to :

$$P(d_j|q) \propto P(d_j \wedge q) \tag{2.10}$$

$$\propto \sum_{\forall \vec{k}} P(d_j \wedge q|\vec{k})P(\vec{k}) \tag{2.11}$$

$$\propto \sum_{\forall \vec{k}} P(d_j|\vec{k})P(q|\vec{k})P(\vec{k}) \tag{2.12}$$

Notice that d_j and q are d-separated given $\vec{k}$ as represented in figure 2.4(b). d_j and q are so mutually independent which allows to write $P(d_j \wedge q|\vec{k}) = P(d_j|\vec{k})P(d_j|\vec{k})$

2.1.2.2 PageRank

PageRank (Brin and Page, 1998) is a link-analysis model for information retrieval that stimulates random navigation of users on the Web. PageRank defines two random walk probabilities. First, the probability that user jumps to a random page with a probability b. Second, the probability that user moves to another page through one link of actual page with a probability $(1 - d)$. A

PageRank score $PR(p_i)$ that estimates the probability of visiting a page p_i is computed iteratively as follows.

$$PR^k(p_i) = \frac{d}{N} + (1-d) \sum_{p_j \in C(p_i)} \frac{PR^{k-1}(p_j)}{|C(p_i)|} \tag{2.13}$$

where $k > 0$ is the iteration number. N is the number of pages in the graph. $C(p_i)$ is the set of predecessors of page p_i. $|C(p_i)|$ is the number of outgoing links on page p_i. d is a random walk parameter. The closer d to 0, the higher importance is given to random jump probability. As d moves towards 1, higher importance is given to the graph structure. For convenience, d is set to 0.15 (Brin and Page, 1998).

Iteration process is continued until converge. This state is obtained if no changes on ranking list is observed for n successive iterations $Rank^k(p_i) = Rank^{k-n}(p_i)$.

Figure 2.5 shows an example of a Web graph with respective PageRank score for each page. PageRank assigns higher scores to a well connected Web page pointed by well connected pages too. The idea behind this principle is to identify authoritative pages in graph.

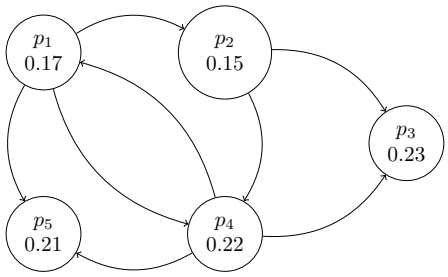


Figure 2.5: PageRank scores of simple Web graph

PageRank estimates the probability of arriving to a particular page according to the random surfer behavior. This probability is assimilated to page authority in the graph. However, PageRank scores are independent from the query topic and do enable selecting relevant documents in response to the query. As a result, PageRank model is integrated with topical-based components in order to select candidate relevant documents. One solution consists of linearly combining topical score $RSV(p_i)$ computed by a classical model with PageRank score $PR(p_i)$. Final document score is given by:

$$P(p_i|q) = \alpha RSV(p_i, q) + (1-\alpha) PR^k(p_i) \tag{2.14}$$

Where $\alpha \in [0, 1]$ is a tuning parameter. The closer α to 1, the highest importance is given to authoritative pages. The closer α to 0, the highest importance is given to document-query similarity.

2.1.3 Retrieval evaluation

Retrieval evaluation provides quantitative metrics for comparing the performances of information retrieval models and studying the impact of involved factors on the retrieval effectiveness. Two categories of retrieval measures are identified: (*i*) recall and precision based metrics evaluates the retrieval effectiveness; (*ii*) rank-oriented measures evaluate the ranking accuracy. We present in what follows the main evaluation measures in information retrieval.

2.1.3.1 Recall and precision

Recall: The Recall evaluates the ability of an information retrieval system to return relevant documents in the answer set. Recall measure is defined by the fraction of retrieved relevant documents over the set of relevant documents in the collection. Let Q be a set of $|Q|$ queries. Recall value is averaged over the set of queries as follows:

$$Recall = \frac{1}{|Q|} \sum_{q_j \in Q} \frac{|S_j \cap R_j|}{|R_i|} \tag{2.15}$$

where S_j is the set of retrieved documents for query q_j. R_j is the set of relevant documents for query q_j.

Precision: The precision measure evaluates the ability of an information retrieval system to return relevant documents in the top of the answer set. The precision is defined as the fraction of relevant documents in the retrieved set. Given a set of queries Q, the precision of an information retrieval system is defined by:

$$Precision = \frac{1}{|Q|} \sum_{q_j \in Q} \frac{|S_j \cap R_j|}{|S_j|} \tag{2.16}$$

Precision at n (P@n): Precision at the n^{th} position $P@n$ computes the precision of a retrieval system over the top n retrieved document. This metric evaluates the system ability to return relevant documents over the top n of results. $P@n$ is mainly used when users are basically interested in the top of retrieved documents, e.g., Web search. Given a set of queries Q, $P@n$ is defined by:

$$P@n = \frac{1}{|Q|} \sum_{q_j \in Q} \frac{|S_{jn} \cap R_j|}{|S_{jn}|} \tag{2.17}$$

With S_{jn} refers to the set of top n retrieved documents for query q_j, ranked by score.

Mean average precision (MAP): Given a set of queries Q, MAP precision averages precision values at each relevant retrieved documents. This measure evaluates the ability of a system to retrieve and return relevant documents in top of retrieved ones. MAP precision is computed as:

$$MAP = \frac{1}{|Q|} \sum_{q_j \in Q} \frac{1}{|R_j|} \sum_{k=1}^{|R_j|} p(R_j[k]) \tag{2.18}$$

Where $R_j[k]$ is the rank of the k^th relevant document in the retrieved set R_j. $p(R_j[k])$ corresponds to precision at $R[k]$ as defined in formula 2.17.

2.1.3.2 Rank-oriented measures

Normalized Discounted Cumulative Gain (NDCG): The Normalized Discounted Cumulative Gain (Järvelin and Kekäläinen, 2002; Kekäläinen, 2005) evaluates the usefulness of a retrieval system for retrieving and ranking documents by decreased order of relevance. In contract of binary relevance judgment used by recall-precision metrics, $NDCG$ supports gradual relevance scale. First, Discounted Cumulative Gain DCG is computed at the n^{th} position for each query $q_j \in Q$ as follows.

$$DCG_j^n = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i} \tag{2.19}$$

Where rel_i is the user relevance assignment to the i^{th} document. The $NDCG@n$ is computed as:

$$NDCG@n = \frac{\sum_{q_j \in Q} DCG_j^n}{\sum_{q_j \in Q} IDCG_j^n} \tag{2.20}$$

Where $IDCG_j^n$ stands for idealized DCG_j^n produced by the perfect ranking of retrieved set where $rel_{i+1} \geq rel_i$.

2.2 Social network analysis

Social network analysis is a network paradigm introduced in social sciences that studies the relationships between individuals (actors) (Parsons, 1949). In contrast of statistical and quantitative network analysis that estimate a probability distribution of true tendency, social network analysis is characterized with a mathematical approach that studies the social network status (Hanneman and Riddle, 2005).

Social network analysis is more a branch of "mathematical" sociology than of "statistical or quantitative analysis", [...] Mathematical approaches [...] tend to regard the measured relationships and relationship strengths as accurately reflecting the "real" or "final" or "equilibrium" status of the network.

Social network analysis focuses on actor's behavior instead of actor's attributes (Pinheiro, 2011). Even though attributes describe actors at personal level, they can not lead to a conclusion about the social interactions. Actors with similar attributes may present different behaviors given the influence of their social neighborhood. Thus, social network analysis investigates social relationships and social network structure as a key feature to understand interactions in the social context.

In this section, we first introduce basic concepts of social networks then we present an overview of centrality measures for social network analysis.

2.2.1 Social networks

A social network is defined by a set of actors who share several relationships with each other (Hanneman and Riddle, 2005). Actors refer mainly to persons but represent, in a boarder context, institutions, communities, information items, etc. A social network may involve one or more types of actors as shown by professional social networks where two types of actors are present: workers and companies. Social relationships refer to the social interactions that involve two or several actors such as friendship and partnership. Both actors and social relationships evolve over the time. New actors and relationships may appear in the social network, others may disappear.

Social networks are represented with a graph $G = (V, E)$ where the set of nodes V represents actors and the set of edges $E = V \times V$ represent relationships between them. In the case of undirected social network, an edge (v_i, v_j) represents symmetric relationships between two actors v_i and v_j. Friendship is a typical example of undirected relationships. In the case of directed social network, an edge (v_i, v_j) represents a directed relationship from v_i to v_j. For instance, email communication is represented by a directed edge (v_i, v_j) where v_i represents the sender and v_j is the recipient. In order to highlight key actors in the social network and indicate the strength of social relationships, representative weights may be assigned to nodes and edges.

Figure 2.6 illustrates a social network of 10 persons. A bidirectional edge denotes a reciprocal relationship between actors. In this example, all social network actors are connected to at least one node in the graph. If we remove nodes p_6, the graph is no longer connected. The new graph includes 2 separated subgraphs, called *"components"*. The most populated component in the graph is known as the *"giant"* component.

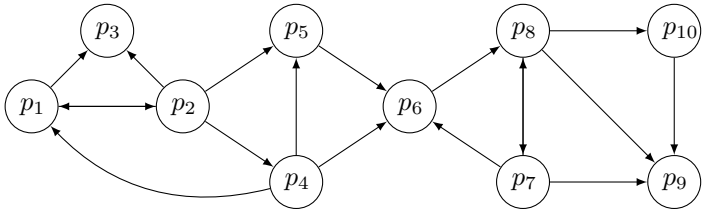


Figure 2.6: Social network example

Social relationships may involve several actors at the same time such as video-conferencing service where many users attend simultaneously the same conference. In this case, the social network is represented by hypergraphs. Furthermore, multiple edges with the same source and destination may be defined between actors sharing several relationships such as friends and colleagues. In the case, the social network is represented by multigraphs.

In order to identify key actors in the social network, namely popular persons, a set of centrality measures are applied on the social network. These measures are discuss in more detail in the following section.

2.2.2 Centrality measures

Centrality measures are structural attributes of social network nodes (Freeman, 1979) that evaluate the importance of actors based on their position in the social network. We detail in what follows, the main centrality measures proposed in the literature.

Degree centrality is a simple centrality measure proposed by Nieminen (1974) that counts the number of adjacent nodes. Degree centrality is defined by:

$$C_D(v_i) = \sum_{\forall v_j \in V} a(v_i, v_j) \tag{2.21}$$

where $a(v_i, v_j) = 1$ if $(v_i, v_j) \in E \vee (v_j, v_0) \in E$, 0 otherwise. A normalized value of Degree centrality in computed as follows.

$$C_D(v_i) = \frac{\sum_{\forall v_j \in V} a(v_i, v_j)}{|V| - 1} \tag{2.22}$$

In the case of directed social networks, two measures of Degree centrality are identified. Indgree Centrality $C_D^-(v_i)$ counts the number of nodes predecessors. Outdgree Centrality$C_D^+(v_i)$ counts the number of node successors.
Degree centrality measures the social activity of a person (Kirchhoff et al., 2008). Higher is the Degree value, the most intensive activity the respective actor shows in the network. Degree centrality is also interpreted as a form of popularity.

Closeness centrality (Sabidussi, 1966) evaluates the proximity for each actor to the rest of the network nodes. Closeness centrality takes into consideration direct connections as well as indirect connections through intermediary nodes. Closeness centrality is defined by:

$$C_C(v_i) = \frac{1}{\sum\limits_{\forall v_j \in V} d(v_i, v_j)} \tag{2.23}$$

where the distance $d(v_i, v_j)$ between nodes v_i and v_j is defined by the length of the shortest path separating the two nodes. A normalized Closeness centrality is defined as follows (Beauchamp, 1965).

$$C_C(v_i) = \frac{|V| - 1}{\sum\limits_{\forall v_j \in V} d(v_i, v_j)} \tag{2.24}$$

Closeness centrality measures the reachability, the reciprocally and the independence of a person to the social network (Kirchhoff et al., 2008). Qualified with a high Closeness centrality, an actor is able to spread an information quickly to a large fraction of the social network.

Betweeness Centrality (Anthonisse, 1971; Freeman, 1977) evaluates the role of an actor in the communication flow between social network nodes. For $v_i \neq v_j \neq v_k$, Betweeness centrality is defined by:

$$C_B(v_i) = \sum_{\forall v_j \in V} \sum_{\forall v_k \in V} b_{jk}(v_i) \tag{2.25}$$

where $b_{jk}(v_i)$ refers to the length of the shortest connecting v_j and v_k through v_i.

Betweeness centrality evaluates the ability of an actor to control the communication flow. It also highlights nodes linking disparate subgraphs in the social network. Betweeness centrality identifies intermediators (*e.g.*, gatekeeper, broker), teams coordinators and interdisciplinary authors (Leydesdorff, 2007).

Eigenvector centrality represents a family of centrality metrics that measures the centrality of a node depending of its neighborhood. Katz centrality (Katz, 1953) and Bonacich's power centrality (Bonacich, 1987) are two examples of eigenvector centrality indicators that evaluate the influence of an actor in the social network. Similarly, PageRank (Brin and Page, 1998) and HITS (Kleinberg, 1999) algorithms identify important nodes in the graph. In particular, PageRank and HITS's Authority score identify network authorities while HITS's Hub score highlights central nodes in the network.

Table 2.1 presents a summary of centrality measures cited above.

	Degree	Close.	Between.	Katz	Bon.	PageR.	HITS	
							Auth.	Hub
Activity	■	□	□	□	□	□	□	□
Popularity	■	□	□	□	□	□	□	□
Gregariousness	■	□	□	□	□	□	□	□
Reachability	□	■	□	□	□	□	□	□
Independence	□	■	□	□	□	□	□	□
Connectedness	□	□	■	□	□	□	□	□
Intermediateness	□	□	■	□	□	□	□	□
Influence	□	□	□	■	■	■	■	■
Authority	□	□	□	□	□	■	■	□
Hubness	□	□	□	□	□	□	□	■

Table 2.1: Proprieties of centrality measures

2.3 Social information retrieval

Social information retrieval is a novel research area that has emerged since early 2000s. It brings together two research fields: information retrieval and social network analysis. Kirsch et al. (2006) defined social information retrieval as the incorporation of social networks data into the information retrieval process.

> *Social information retrieval systems are distinguished from other types of information retrieval systems by the incorporation of information about social networks and relationships into the information retrieval process Kirsch et al. (2006).*

In fact, the emergence of the social Web and the significant position that users has acquired in information producing and consuming processes has challenged traditional information retrieval approaches, being focused on document level regardless of the surrounding social context. To tackle this issue, social information retrieval provides a new generation of retrieval models that exploit the social network structure and data to enhance the retrieval process.

We present in this section the main social approaches for information access and retrieval. Afterward, we focus on social network models and relevance factors. Finally, we give an insight view into search and ranking processes within social networks.

2.3.1 Social content graph

Social data (*e.g.*, documents, comments, annotations and ratings, *etc*) as well as the mutual interactions between persons and content represent the collective intelligence mechanism that people has developed on Web. Thanks to the *" wisdom of the crowd"*, people are able to attain more accurate results rather than individual efforts (Surowiecki, 2005). In fact, person and content interac-

tions can be exploited to enhance the user experience on the Web typically for information search and access.

In contrast of the traditional Web where Web pages are the central and the only type of entities, persons and content play, side-by-side, a key role in the social content graph. As shown in figure 2.7, the social content graph includes 4 types of interactions: content-to-content, content-to-person, person-to-person and person-to-content. These interactions define the social producing context and the social consuming content of information.

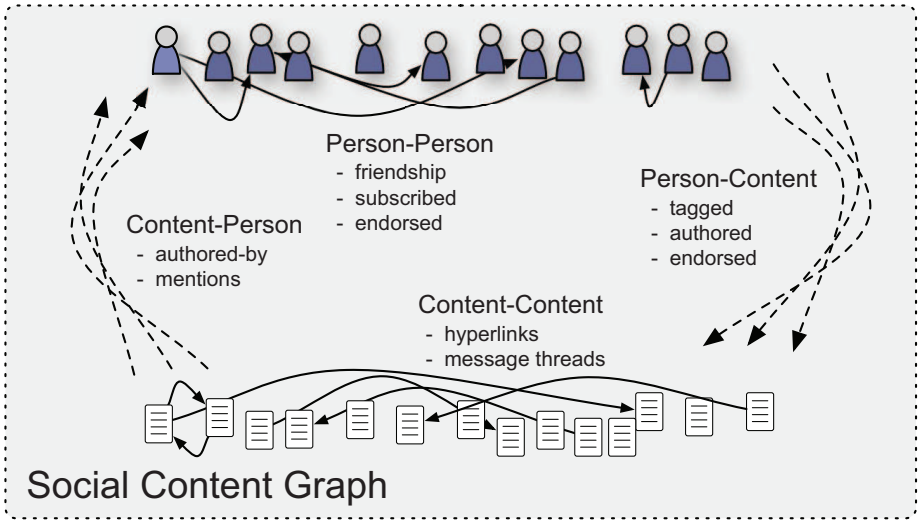


Figure 2.7: The Social Content Graph (Amer-Yahia et al., 2007)

Social information retrieval systems exploit content-to-content interactions such as hyperlinks in order to highlight central documents in the Web graph. Retrieval results are therefore ranked by their centrality. Content-to-Person interactions such as mentions enable to identify focused persons that represent the main topic of the content. For instance, user searching for the biography of *Charlie Chaplin* may be interested in some texts where this famous actor is mentioned. Person-to-person interactions help to identify central persons in the social network typically experts within a query topic. Finally, Person-to content such as tags and comments may reflect user interest on published content and stands as potential relevance feedback for his queries.

The topology of social content graph may differ according to the social application but include in a border context two main types of entities: actors and data. Actors represent persons. They may have different roles in the social graph. On the other hand, data refer to information items. Social relationships between actors and data could be explicit or implicitly extracted from social interactions.

For instance, Wikipedia social content graph includes 2 connected networks as shown in figure 2.8 (Korfiatis et al., 2006). Article network is comprised of Wikipedia articles. Edges in this layer represent hyperlinks. Contributor network is comprised of Wikipedia contributors. Two contributors are connected if they collaborated together in the same project. Edges connecting contributor to an article indicate if this person has participated to write respective article.

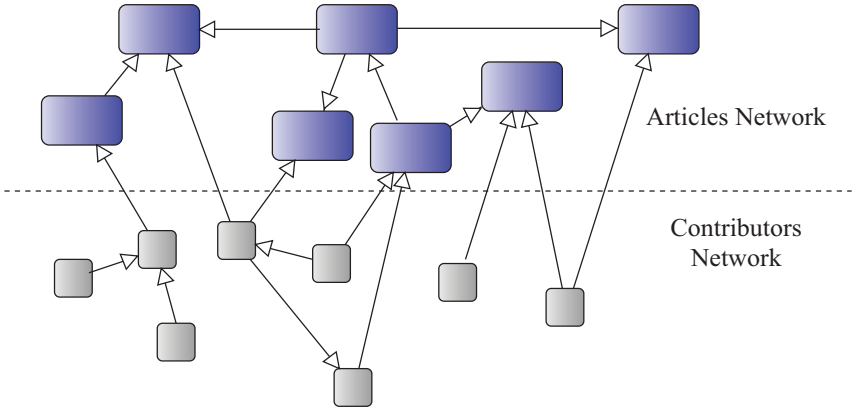


Figure 2.8: Network layers in the wiki publication model (Korfiatis et al., 2006)

The social network of Wikipedia is formally represented with a graph $G := (V, E)$. The set of nodes $V = A \cup C$ includes articles A and contributors C. The set of edges $E = V \times V$ denotes the relationships between graph nodes. $A \times A$, $C \times C$ and $C \times A$ correspond to the subset of hyperlinks, collaborations and authorship links, respectively.

2.3.2 Social retrieval processes

Social information retrieval differs to other retrieval approaches by integrating the social network structure into the retrieval processes. The retrieval cycle is conducted over three elementary processes: social network extraction, social network analysis and document relevance ranking (Kirchhoff et al., 2008). These processes are implemented as an extension to traditional information retrieval systems as shown in Figure 2.9.

Social network extraction. In the first step, the social network structure is extracted from the collection of documents. In spite of explicit social relationships, new social relationships are derived from content and metadata. For example, co-citations are extracted from citation graph. This step is performed independently from user queries.

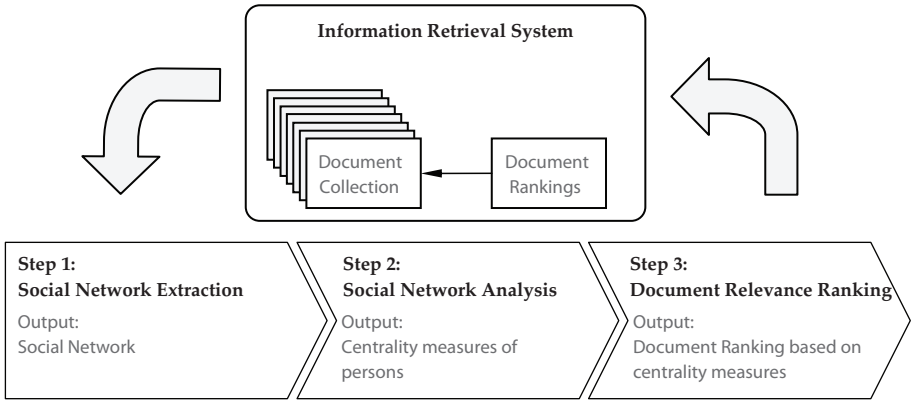


Figure 2.9: Principle model of the social network enhanced information retrieval system (Kirchhoff et al., 2008)

Social network analysis. In the second step, social network analysis methods are applied on the social network in order to identify key entities. A social relevance score is assigned to each network actor based on centrality measures introduced in section 2.2.2.

Document relevance ranking. In the third step, a query-based relevance score is combined with social-based relevance score in order to produce a final ranking of documents. Document that stratify user information need and with related actors are central in the social network are ranked at the top of the results. Query-based relevance score and social-based scores are combined either by an integrated approach or a modular approach (Amer-Yahia et al., 2007).

An Integrated approach represents relevance factors as transition probabilities on the social content graph then applies a random walk algorithm such as PageRank in order to rank retrieved results. In particular, a query-based score is assigned as an initial score to document nodes. Additional relevance factors such as document similarly and author expertise are represented through edge weights. Afterward, nodes scores are propagated through the network edges in order to identify central documents in the social network.

A Modular approach computes an independent score for each relevance feature then estimates a final score using an aggregation function. In particular, the final score is computed as the sum of query-based relevance score and a social-based relevance.

$$score(q, d) = RSV(q, d) + S_d \tag{2.26}$$

Where $RSV(q, d)$ is document-query similarity and S_d is the social score of document d computed as the centrality of respective authors in the social network.

2.3.3 Social retrieval tasks

Novel social approaches come into sight to satisfy user's information needs initiated by the new social practices on the Web. Table 2.2 classifies these approaches into four main categories according to the retrieval tasks:

Recommendation: Unlike collaborative recommendation and filtering approaches that extract recommendation rules from implicit associations between persons and documents, social recommendation approaches incorporate explicit social relationships in the social content graph in order to present trusted recommendations. These approaches propose generalized models for items' recommendation (Guy et al., 2010) as well as specific recommendation models for personalized access (Baluja et al., 2008), product suggestion (Ma et al., 2009; Jamali and Ester, 2009) and tag prediction (Heymann et al., 2008b).

Social information extraction: User needs to know more about his social network structure for better understanding and exploring the social graph. Thus, social retrieval approaches have investigated the social properties of the content graph and propose several mining models that make available new knowledge from the social network (Domingos, 2005; Tang et al., 2008; Schifanella et al., 2010; Adamic et al., 2008; Zhang, 2011). For instance, community clustering approaches are proposed in this context to help users finding persons with similar interest in the social network (Newman and Girvan, 2004; Girvan and Newman, 2002; Crandall et al., 2010). Moreover, collaboration analysis methods enable users to conduct promising collaboration with complement actors in the social network (Bird et al., 2006; Welser et al., 2011; Kumar et al., 2010).

Opinion retrieval: Blogs, comments, reviews and ratings are different kinds of opinionated content on the Web that allow people to express their opinion about events, products and services (Song et al., 2007; Bodendorf and Kaiser, 2009). Such information allows other persons to learn from similar experience and then make accurate decision. Before booking a hotel room, for instance, users would like to check customer comments about room services. The challenges of opinion retrieval approaches, is to detect opinionated content and determinate associated sentiment (*e.g.*, negative, neutral or positive sentiment). Social network properties allows on the first hand to detect public and community sentiments (Pang and Lee, 2008; Hui and Gregory, 2010; Danescu-Niculescu-Mizil et al., 2009; Jansen et al., 2009), and on the other, to track opinion influence in the social graph (O'Connor et al., 2010; Thelwall et al., 2011; Wu and Huberman, 2004).

People search: Searching the social network allows users to find other persons in the social network that satisfy some criterion or show particular social properties (Adamic and Adar, 2005). In case of professional social

Task	*Interest*	*Social networks*	*Literature*
Recommendation	Item recommendation, media recommendation, tag prediction, product suggestion, relationship suggestion.	tagging, trust network, media sharing, movies, products, news, travel etc.	(Heymann et al., 2008b; Ma et al., 2009; Guy et al., 2010; Baluja et al., 2008; Jamali and Ester, 2009)
Social information extraction	Collaboration patterns, community structure, link prediction, network evolution, knowledge sharing, innovation network.	Wiki, professional network, scientific network, question answering, collaborative network, etc.	(Bird et al., 2006; Tang et al., 2008; Domingos, 2005; Newman and Girvan, 2004; Girvan and Newman, 2002; Schifanella et al., 2010; Crandall et al., 2010; Adamic et al., 2008; Welser et al., 2011; Kumar et al., 2010; Zhang, 2011)
Opinion retrieval	Sentiment analysis, opinion search, opinion polarization, leadership, trends.	Blogs, microblogs, forums, customers, comments, rating, news, etc.	(Song et al., 2007; Pang and Lee, 2008; Bodendorf and Kaiser, 2009; Hui and Gregory, 2010; O'Connor et al., 2010; Thelwall et al., 2011; Wu and Huberman, 2004; Danescu-Niculescu-Mizil et al., 2009; Jansen et al., 2009)
People search	People and expert search, identifying key actors, popularity, authority, influence, trust	Email, blogs, personal network, professional network, scientific network, etc.	(Deng et al., 2008; Macdonald and Ounis, 2008; Balog et al., 2008; Pal and Counts, 2011; Fu et al., 2007; Kazai and Milic-Frayling, 2008; Adamic and Adar, 2005)
Social search	Web search, trust search, social rank, social network search, collaborative search, query expansion, recency ranking.	Bookmarking, media sharing, log data, wiki, blogs, microblogs, news, personal network, etc.	(Korfiatis et al., 2006; Dong et al., 2010; Bao et al., 2007; Krause et al., 2008; Kirchhoff et al., 2008; Efron, 2011; Mislove et al., 2006; Ma et al., 2008; Kirsch et al., 2006)

Table 2.2: Retrieval tasks in social content graph

network, users need to identify people occupying particular positions in some companies (Macdonald and Ounis, 2008). Similarly, scientific express their need to identify authoritative researchers in their research area (Deng et al., 2008). Other properties are investigated in respect of the social networking purpose such as popularity (Kazai and Milic-Frayling, 2008), influence (Pal and Counts, 2011) and expertise (Balog et al., 2008; Fu et al., 2007). People search approaches proposes a verity of social network search models based on both personal profile and social relationships.

Social search: In addition to classic documents, users express their information need to search for social generated content. In this aim, social search approaches ensure retrieval task within the social content graph while taking into account the structure of the social network (Korfiatis et al., 2006). Accordingly, documents are ranked by their relevance as well as their significance in the social network (Kirchhoff et al., 2008; Kirsch et al., 2006). In addition to classic documents (Bao et al., 2007; Krause et al., 2008), search may target different types of data such as news articles (Dong et al., 2010) and blogs (Efron, 2011; Ma et al., 2008). Social search approaches apply specific retrieval features for each type of data in respect of corresponding social network properties .

Conclusion

We introduced in this chapter basic concepts of information retrieval and main state-of-the-art retrieval models proposed for this aim. Moreover, we gave a brief introduction of social network analysis methods and we discussed principal centrality measures that evaluate the importance of social actors. Finally, we gave an overview of social information retrieval field which in fact a multidisciplinary research fields that bridges information retrieval and social network analysis.

Among the social information retrieval tasks discussed in the end of this chapter, we are particularly interested in social search task. This task exploits the social network structure in order to enhance the retrieval and the ranking processes of information retrieval systems. In the two next chapters, we will focus on this issue, particularly for literature access and microblogging, and we will discuss main social information retrieval approaches proposed for these application domains.

Part I

On using social networks to enhance information access and retrieval: focus on literature access and microblog search

Chapter 3

Applying social networks for literature access

Introduction

Literature access takes advantage of information retrieval approaches in order to provide scientific researchers with useful publications in their research domain (Bollacker et al., 2000). Literature access is in fact an information retrieval task conducted over a collection of bibliographic resources. Since 1990's, several work are proposed in this context for indexing, retrieving and ranking scientific publications (Giles et al., 1998; Schatz et al., 1996; Cleveland and Dataflow, 1998; Humphrey, 1992).

Interestingly, scientific publications have been addressed by several research fields as a conceptual and an evaluation framework due to data quality and richness. In particular, scientific publications were investigated by scientometrics and bibliometrics in order to evaluate the scientific impact of research (Garfield, 1964; Pinski and Narin, 1976; Hirsch, 2005; Cabanac and Hartley, 2013). In addition, scientific publications were used as an evolution framework for information retrieval systems. Popular algorithms that have marked Web search such as PageRank (Page et al., 1999) was primary evaluated over a collection of scientific publications. Besides, scientific publications are considered as promising benchmark for social network analysis. Several social network methods are proposed in this context based on the social network of scientific authors, typically the work of Newman (Newman, 2005; Newman and Girvan, 2004; Newman, 2001b,a).

With the emergence of social Web, literature is again investigated as a tool for designing and evaluating social information retrieval (Kirchhoff et al., 2008; Kirsch et al., 2006). This choice is argued by the fact that scientist publications

as well as scholarly bookmarking services provide valuable data that supports at the same information retrieval as well as social network analysis methods. Such category of social network based approaches of literature access exploits social network data in order to enhance an information retrieval process. Several models for literature recommendation (Farooq et al., 2007), expert search (Deng et al., 2012, 2008) and bibliographic retrieval (Kirchhoff et al., 2008; Kirsch et al., 2006) are proposed in this context.

We discuss in this chapter social information retrieval approaches for literature access that particularly exploit the social network in order to enhance information access and retrieval. For this aim, we first describe the structure of scientific social network and scholarly social bookmarking networks. Moreover, we discuss the main scientific impact measures, in particular social-based metrics for ranking authors. Afterward, we focus on social retrieval and ranking approaches in digital libraries. Finally, we review some examples of social retrieval and access systems proposed by the main academic search engines in the Web.

3.1 Scientific social networks

Authors and articles represent key entities in a scientific publication process. As shown in figure 3.1, these entities are connected with two basic relationships: authorship and citation. Both relationships are directed and explicitly mentioned in article contents.

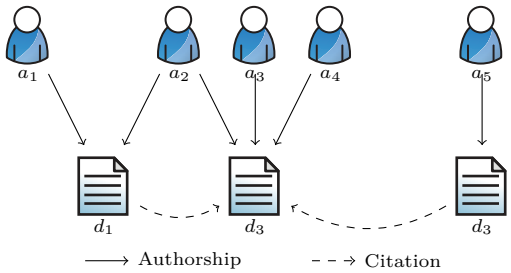


Figure 3.1: Scientific publication network

In order to study researcher's status at the individual level as well as the overall structure of the research community, we need to extract the social networks of scientific authors. In contract of publication network, scientific social networks focus on the social context behind the production of bibliographic resources and analyze author interactions. We present in what follows the main social network models for scientific research, namely co-authorship network and citation network.

3.1.1 Co-authorship network

Co-authorship network is a widely used social network representing collaboration among researchers (Bordons and Gómez, 2000). It associates together authors who have collaborated to produce a scientific publication. Besides collaboration aspects, co-authorship network represents personal acquaintances and shared interest between researchers.

Actually, co-authorship network is built based on co-authoring associations between authors. Since co-authorship is not explicitly mentioned in a publication network, this relationship is inferred from authoring associations that connect authors and articles.

Definition 1 *Let $D = \{d_1, d_2, \ldots, d_n\}$ be a set of articles and $A = \{a_1, a_2, \ldots, a_m\}$ be a set of authors. We define:*

$$\delta_i^k = \begin{cases} 1, & \text{if } a_i \text{ has authored } d_k \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

Two authors a_i and a_j are called "co-authors" if it exists at least one document d_k authored by both authors a_i and a_j.

$$\delta_i^k \delta_j^k = 1 \tag{3.2}$$

3.1.1.1 Network topology

Different approaches are proposed in the literature in order to model co-authorship networks. Basic network model represents authors with a binary and undirected graph.

Definition 2 *Co-authorship network is an undirected graph $G := (V, E)$ where the set of vertices V represents authors and the set of edges $E = V \times V$ denotes co-authorships between them. An edge $(a_i, a_j) \in E$ is defined between each couple of co-authors a_i and a_j.*

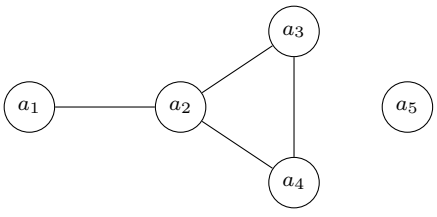


Figure 3.2: Co-authorship network

Figure 3.2 illustrates a co-authorship network extracted from the scientific publication network in figure 3.1.

Undirected models of co-authorship networks show some limits notably the lack of endorsement between authors. Moreover, some social network analysis methods (*e.g.*, PageRank) could not be applied on undirected graphs. To resolve this problem, an undirected co-authorship network is transformed into a directed graph where each undirected edge (a_i, a_j) is replaced with two symmetric and directed edges (a_i, a_j) and (a_j, a_i).

3.1.1.2 Co-authorship weights

Binary co-authorship does not distinguish substantial co-authorships in the authors' scoial network. In fact, occasional co-authorships do not have the same significance as frequent co-authorship where co-authors may conduct an extensive collaboration. Moreover, papers with a few co-author number ensure higher percentage of individual contribution contrary to papers that involve a considerable number of authors. To tackle this issue, Newman (2001a) proposes to weight network edges in respect of co-authorship frequency and the number of co-authors by paper. The strength of collaboration w_{ij} between two authors a_i and a_j is defined by:

$$w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1} \tag{3.3}$$

where n_k is the number of authors for paper d_k.

Notice that collaboration strength is computed equally for each couple of co-authors regardless of edge direction. Nevertheless, co-authorship does not have symmetric significance since one of the involved co-authors may publish more papers with a particular author than anyone else. Liu et al. (2005a) propose so a normalized co-authorship weight that reflects exclusivity in co-authorship activities.

$$\overline{w}_{ij} = \frac{w_{ij}}{\sum_v w_{iv}} \tag{3.4}$$

3.1.2 Author citation network

Author citation network is a social network that represents scientific endorsement between researchers. In contrast of document citation network, author citation network investigates reference relationships between authors instead of documents. Accordingly, author citation network determine influence and knowledge transfer within the scientific community.

Author citation network is built based on author-to-author citation. However, this relationship is not immediately available in the scientific publication graph. Author citation is therefore inferred from document-to-document citations.

Definition 3 *Let d_k and d_s be two documents in the publication graph. Document citation is defined by:*

$$\sigma_k^s = \begin{cases} 1, & \text{if } d_k \text{ cites } d_s \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

A citation relationship is then defined from authors a_i to authors a_j if it exists a document authored by a_i that cites a document authored by a_j.

$$\mathcal{C}_i^j = \begin{cases} 1, & \text{if } \exists\ d_k, d_s \in D : \delta_i^k \delta_j^s \sigma_k^s = 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

We note that citation extraction is confronted to some challenges. This is due to spelling irregularities and name similarity between authors. In addition, citation practices are sometimes ambiguous such in-text citations that correspond to no entry in reference list. To resolve this problem, machine learning techniques are used to extract metadata from bibliographic citations (Lawrence et al., 1999b; Han et al., 2003).

3.1.2.1 Network topology

Author citation network is generated by projecting of document citation links on corresponding authors. These relationships are reproduced for each couple of citing and cited authors according to the direction of document citation. Based on graph formalism, an author citation network represents authors with a directed graph.

Definition 4 *Author citation network is a directed graph $G := (V, E)$ where the set of vertices V represents authors and the set of edges $E = V \times V$ denotes citation between them. An edge $(a_i, a_j) \in E$ is defined from author a_i to author a_j if a_i cites a_j, which corresponds to $\mathcal{C}_i^j = 1$.*

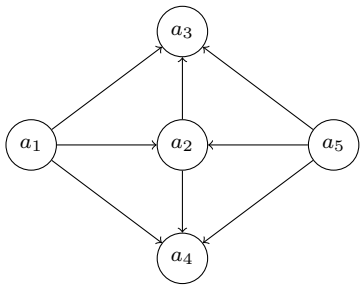


Figure 3.3: Citation network

Figure 3.3 illustrates an example of citation network extracted from the scientific publication network in figure 3.1. In contrast of co-authorship network

presented in 3.2 where authors a_5 is represented with an isolated node, citation network shows higher connectivity between authors. In fact, nodes of this citation network example are represented within a single component where each author is involved at least in one citation relationship.

We note that density of author citation network depends on the size of the citation repository. In fact, citation relationships have more chance to be recognized if document repository covers a large number of journals, conferences and digital libraries. It's also important to build a citation network from recent repository snapshots as the topology of citation network evolves over time (Hummon and Dereian, 1989; Baumgartner and Pieters, 2003). While recent documents take enough time to be known and then to be cited, new relationships appear in the citation graph. At the same time, authors of aged papers keep acquiring new citations.

Having the ability to cite one of his previous papers, authors may generate more incoming citations. This practice has been debated in the literature. While some work propose to remove this type of citations for accurate analysis (Lawani, 1982; Schreiber, 2007; Życzkowski, 2010), other studies argue such practices as self-citation may have different motivations than other citations (Hyland, 2003; Pichappan and Sarasvady, 2002; Phelan, 1999).

3.1.2.2 Citation weights

Each represented author in the citation network inherits incoming and outgoing citations from corresponding papers. Without appropriate weights on network edges, all citations are treated equally regardless of their significance (Diamond Jr, 1986).

The problem of citation weighting is widely discussed in the case of multiple authorships. In fact, received credits should be distributed on co-authors proportionally to their levels of contribution. This idea is not supported by the traditional *"normal count"* scheme where equal citation weights are assigned for each co-author. Assuming that major contribution is made by the first co-author, Cole and Cole (1981) propose a *"straight count"* weighting where received credits are exclusively attributed to the first author. However, this assumption is not always warranted typically when co-authors are listed by alphabetical order. To handle this issue, *"adjusted count"* scheme (Lindsey, 1980; De Solla Price, 1963) assigns received credits proportionally to the number of co-authors. Afterward, weight of author-to-author citation edge is computed as the sum of all references weight.

In order to highlight author productivity, Radicchi et al. (2009) propose to extend *adjusted count* weighting by considering both citing and cited co-author

number. Citation weight is therefore computed as follows.

$$w_{ij} = \sum_{k,s} \frac{\delta_i^k \delta_j^s \sigma_k^s}{n_k n_s} \tag{3.7}$$

where n_k is the number of co-authors for citing paper and n_s is the number of co-author for cited paper.

3.1.3 Co-citation and affiliation networks

In addition to co-authorship and citation networks, authors of scientific publications are represented with different social network models namely co-citation and affiliation networks. Co-citation network (Small, 1973; Gmür, 2003; Ding et al., 2009) connects via a co-citation relationship each couple authors who have been cited by the same publication. This relationship expresses the semantic similarity between connected authors. Affiliation network (Wasserman and Faust, 1994; Faust, 1997; Singh and Getoor, 2007) connects each couple of authors who belong to the same research institution or have participated in the same research event. This type of social networks confirms the social acquaintance between authors.

Besides these social network structures that represent mainly authors of scientific papers, new types of networks could be inferred from scholarly social bookmarking services. The social bookmarking network allows representing the social consuming context of scientific publications in spite of previously cited models addressing the social producing context. Next, we will discuss in more detail the social bookmarking network.

3.2 Scholarly social bookmarking

Social bookmarking is an online and collaborative annotation service that allows users to share and organize a set of items. Unlike Web browser bookmarks and Web directories, social bookmarking services provide users with a free access to their selected resources as well as public resources shared by other users (Yanbe et al., 2007). At the user level, social bookmarks help to reach an already visited item. User bookmarks are also useful at the community level as it enable other members to discover valuable items (Heymann et al., 2008b).

Social bookmarking has especially gained popularity since the lunch of Delicious[1] in 2003. This Web site stood as the leader of social bookmarking services for many years. Similar services for scholarly and academic purposes are subse-

[1] http://delicious.com/

quently introduced, namely, CiteULike[2] (2004), Connotea[3] (2004), BibSonomy[4] (2006) and Mendeley[5] (2008). Scholarly social bookmarking services enable users to share, store, and organize information about scholarly papers (Farooq et al., 2007). In comparison to other social bookmarking services, such domain-specific bookmarking services offer additional features that help researchers to accomplish their tasks such as the generation of BibTeX records. Furthermore, scholarly social bookmarking services cover various research interests unlike digital libraries that target a particular research fields or scientific conferences.

We describe in what follows the structure of social bookmarking networks than we discuss social-based models for literature access that exploit annotators social networks to recommend scientific publications.

3.2.1 Tagging application

By bookmarking an object, user autocratically share this item with other users and add his own annotation to it. Annotations are in fact a set of terms where each entry is represented by a tag. Heymann et al. (2008b) define a social bookmarking system as a set of triples that involves users, tags and objects.

Definition 5 *A social bookmarking system consists of users $u \in U$, tags $t \in T$, and objects $o \in O$. A post is made up of one or more triples (t_i, u_j, o_k) where user u_j annotates object o_k with tag t_i. User can assign tag t_i only one time to the same object o_k.*

For each object o, the set of tags that positively describes an object o is noted R_p. Conversely, R_n is the set of tags that negatively describe object o. R_a represents the set of tags which are actually used to annotate object o. Only set R_a is accessible as R_p and R_n may include other unknown tags not already used.

In the case of scholarly social bookmarking, objects are represented by scientific papers. Figure 3.4 illustrates an example of CiteULike tagging application made of 3 posts: $(Tag\ A, User\ 1, Paper\ ABC)$, $(Tag\ A, User\ 1, Paper\ XYZ)$ and $(Tag\ B, User\ 2, Paper\ ABC)$. In this example the set of tags that annotate $Paper\ ABC$ is defined by $R_a = \{Tag\ A, Tag\ B\}$. The set of tags that annotate $Paper\ XYZ$ is $R_a = \{Tag\ A\}$.

Tags are mainly used to describe the content of scientific papers. In contrast of keywords, tags show a high level of abstraction that reflect people's understating of the document (Li et al., 2008). Another aim of tags is to describe document context and attributes such as conference name (*e.g.*, *"SIGIR"*, *"WIC'12"*) and paper's type (*e.g.*, *"poster"*, *"survey"*). In the same cases, tags have a personal

[2]http://www.citeulike.org/
[3]http://www.connotea.org/
[4]http://www.bibsonomy.org/
[5]http://www.mendeley.com/

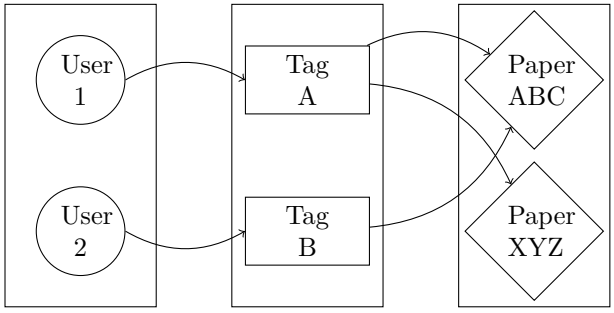


Figure 3.4: Anatomy of tag application in CiteUlike (Farooq et al., 2007)

purpose or target a specific audience such organizational tags (e.g., *"to read"*, *"to review"*) and group tags (e.g., *"IAR Group"*, *"day 3"*).

Users of social bookmarking services are connected with friendship relations. This social relationship distinguishes social bookmarking applications from classical bookmarking tools. The totality of these relations of friendship forms the social network of taggers. It reflects the user interest and the mutual trust between each others. We note that user friendship is either mentioned explicitly (*e.g.*, Mendeley) or implicitly established through groups subscription to (*e.g.*, CiteULike). Other works propose to extract friendship from common tags and documents (Li et al., 2008; Symeonidis et al., 2008; Santos-Neto et al., 2009). The similarity between two users connected with friendship association is computed based on the tags overlap (Schenkel et al., 2008; Amer-Yahia et al., 2008).

$$O(u, u') = \frac{2 \times |tagset(u \wedge u')|}{|tagset(u)| + |tagset(u')|} \tag{3.8}$$

We note that friendship similarly can be computed using social network metrics such as the social distance between users in the graph.

3.2.2 Literature recommendation

The collaborative intend of social bookmarking services and the role of tags in describing both bookmark content and user preferences have made from bookmaking data a natural resource to generate effective recommendations (Zhang et al., 2011). A vast amount of works in different application domains have investigated social bookmarking data in order to enhance the recommendation process. Some of the representative works are listed next.

- Web search (Wu et al., 2006; Yanbe et al., 2007; Jäschke et al., 2007; Bischoff et al., 2008; Heymann et al., 2008b).
- E-commerce (Linden et al., 2003; Givon and Lavrenko, 2009; Sen et al., 2009).

– Media sharing (Geisler and Burns, 2007; Sigurbjörnsson and van Zwol, 2008; Garg and Weber, 2008; Liu et al., 2010; Larson et al., 2011).

Besides these application domains, several approaches that use scholarly social bookmarking data have addressed literature recommendation in scientific research (Parra and Brusilovsky, 2009; Bogers and van den Bosch, 2008; Guan et al., 2010; Jomsri et al., 2011). The approaches proposed innovative models for tag suggestion and scientific papers recommendation. In contrast of the first proposed approaches for literature access that investigates mainly paper content and citation graph (Yarowsky and Florian, 1999; Basu et al., 2001; Dumais and Nielsen, 1992; McNee et al., 2002; Strohman et al., 2007), social bookmarking based approaches have mainly addressed literature recommendation task from the user point of view.

Recommendation models for literature access are classified into two main categories: Collaborative filtering based approaches and graph based approaches.

3.2.2.1 Collaborative filtering based approaches

Collaborative filtering based approaches propose to recommend similar papers that present the same tags (*item based filtering*) or bookmarked by the same users (*user based filtering*). Experimental studies on CiteULike dataset show that user based filtering presents better recommendation results compared to item based filtering in the context of scholarly bookmarking system (Bogers and van den Bosch, 2008).

Parra and Brusilovsky (2009) propose a neighbor-weighted collaborative filtering algorithm that extends classic collaborative filtering models (Schafer et al., 2007). First, a set of neighbors $\mathcal{N}_u$ is selected for user u requesting recommendation. $\mathcal{N}_u$ includes all the users sharing at least one common article or tags with user u. The similarity between users $Sim(u, n)$ is then computed using *Pearson's correlation* coefficient (Rodgers and Nicewander, 1988). A primary recommendation score is attributed to each document based on user rates r as follows (Schafer et al., 2007).

$$pred(u, i) = \bar{r}_u + \frac{\sum\limits_{n \in \mathcal{N}_u} Sim(u, n) \times (r_{ni} - \bar{r}_n)}{\sum\limits_{n \in \mathcal{N}_u} Sim(u, n)} \tag{3.9}$$

Paper score is finally computed by incorporating raters count nbr_i to ensure collective endorsement of the paper.

$$pred'(u, i) = log_{10}(1 + nbr_i) \times perd(u, i) \tag{3.10}$$

3.2.2.2 Graph based approaches

Graph based recommendation approaches explore the structure of the social bookmarking network in order to recommend central resources in the graph. Based on the idea that *"a resource which is tagged with important tags by important users becomes important itself"*, Jäschke et al. (2007) propose adapted version of FolkRank algorithm for tag suggestion within their publication-sharing system BibSonomy[6]. FolkRank is a PageRank-like algorithm for search and ranking over folksonomies previously proposed by authors (Hotho et al., 2006).

A folksonomy is molded by tripartite and undirected hypergraph where $G_F = (V, E)$ where $V = U \cup T \cup R$ is the set nodes including users U, tags T and resources R. $E = U \times T \times R$ is the set of hyperedges that connects nodes from different types. In order to apply FolkRank algorithm, hypergraph G_F is transformed into a directed graph where edges are weighted according to the co-occurrences of each pair of users, tags and resources.

The weight of nodes is computed iteratively using the following formula:

$$\vec{w} \leftarrow dA\vec{w} + (1-d)\vec{p} \tag{3.11}$$

where $\vec{w}$ is the weighting vector. A is the row-stochastic adjacency matrix of G_F, $\vec{p}$ is the topic preference vector and $d \in [0, 1]$ is a random surfer parameter as defined in PageRank.

Given user u and resource r, the set of n candidate tags $\tilde{T}(u, r)$ is defined by:

$$\tilde{T}(u,r) := \arg\max_{t \in T}^{n} \sum_{v \in \mathcal{N}_u} Sim(\vec{x}_u, \vec{x}_n)\delta(v,t,r) \tag{3.12}$$

where $\mathcal{N}_u$ is the set of user u neighborhoods. X is the binary matrix of user neighborhood extracted from either common resources or common tags. $\delta(v, t, r) = 1$ if user v tagged a resource r with tag t, 0 otherwise.

Finally, the set $\tilde{T}(u, r)$ is ranked according to the corresponding weights in $\vec{w}$ to generate to the top-n tag recommendations for a user's resource.

3.3 Scientific impact analysis in bibliographic network

The scientific impact refers to the set of quantitative and qualitative metrics that evaluate scientific contributions. Although it is interpreted as the number of received citations, the scientific impact is a general concept that may express various aspects of research activities (Bollen et al., 2009).

[6] http://www.bibsonomy.org/

Different impact metrics are proposed in the literature in order to measure the scientific quality of journals, papers and researchers. We discuss in this section main impact metrics, typically the graph-based and social network based metrics.

3.3.1 Impact of scientific publications

The problem of evaluating the scientific quality of publications has been addressed in 1960s by the bibliometric research field. In this context, Garfield (1964) proposed the Impact Factor (IF) that measures the quality of scientific journals by computing, for all published papers in the journal, the average number of received citation during the last two years. For a single paper, the citation Impact counts the number of its received citations. Recursive Impact Factor (Pinski and Narin, 1976) was proposed later in order to highlight citations in higher impact journals.

Recent works have investigated the structure of citation network in order to evaluate the prestige of publications. For instance, Journal PageRank algorithm measures journal impact by applying a weighted version of PageRank on journal citation network (Bollen et al., 2006). Journal score is defined by:

$$PR_w(v_i) = \frac{1-\lambda}{N} + \lambda \sum_j PR_w(v_j) \times w(v_j, v_i) \tag{3.13}$$

with propagation weight $w(v_j, v_i)$ is defined by the proportion of citation links from journal v_j to v_i over the sum of outgoing citation of journal v_j.

Likewise, several works (Ma et al., 2008; Li and Willett, 2009) propose to apply PageRank algorithm on paper citation network to evaluate the scientific impact of papers. For instance, *ArticleRank* Li and Willett (2009) computes an authority of articles as defined next:

$$ArticleRank(A) = 1 - d + d \times \sum_{i=1}^{n} \frac{ArticleRank(P_i)}{NR(P_i)} \tag{3.14}$$

with $NR(P_i)$ is the number of citation from paper P_i.

Also based on PageRank algorithm, *CiteRank* (Walker et al., 2007) computes a time-ware authority score of each paper in citation network. In particular, the probability of jumping to a paper i is computed proportionally its recency as defined in next formula.

$$\rho_i = e^{-age_i/\tau_{dir}} \tag{3.15}$$

where age_i is the age of paper i and τ_{dir} is the average age of the initial paper.

Somehow similar, *FutureRank* (Sayyadi and Getoor, 2009) estimates for each paper an expected number of citation that will obtained in future. In addition

to expected citation (R^{Time}), *FutureRank* takes into account the *PageRank* score of articles in citation network ($M^c R^c$) and HITS's authority score in authorship network .

$$R^P = \alpha M^c R^c + \beta M^{A^T} R^A + \gamma R^{Time} + (1 - \alpha - \beta - \gamma)\frac{1}{n} \tag{3.16}$$

with $\alpha + \beta + \gamma + (1 - \alpha - \beta - \gamma) = 1$ and n is the number of papers.

3.3.2 Author impact metrics

Similarly to impact metrics for journals and papers, author impact metrics evaluate the scientific productivity of an author and estimate his influence on the research community. The h-index (Hirsch, 2005), which is one of the first metrics proposed for this aim, measures the impact of an author based on the distribution of received citations. Given a set of published papers, *"a scientist has index h if h of his/her N_p papers have at least h citations each"*. Inspired by this metric, the g-index (Egghe, 2006) computes author's impact as the *"the (unique) largest number such that the top g articles received (together) at least g^2 citations"*, where the set of articles are ranked by decreasing order of citations. A simple impact metric is proposed by Google in 2011, namely the *i10*-index[7], measures the impact of an author as the number of papers with at least ten received citations.

Another category of approaches, propose to evaluate the impact of authors by applying centrality measures on the social network of authors, typically co-authorship (Mutschke, 2003; Yin et al., 2006; Vidgen et al., 2007; Liu et al., 2007). Different measures have been investigated in these works including Degree, Betweenness, Closeness and Eigenvector metrics. An empirical study on 16 journals in the field of library and information science (Yan and Ding, 2009), shows that most of centrality measures significantly correlate with citation counts. Some of these measures have been adapted in order to evaluate the scientific impact of authors (Newman, 2005; Brandes, 2008).

A wide range of PageRank-like algorithms are proposed in the literature in order to study author's impact. For instance, AuthorRank algorithm (Liu et al., 2005a) computed on weighted co-authorship network define the researcher impact to the authority in the social network. Original PageRank weighting $\frac{1}{|C(p_j)|}$ in formula 2.13 is replaced with a co-authorship weight $w_{i,j}$ as defined in equation 3.8. The impact of an author is then computed as follows:

$$AR(i) = (1 - d) + d \sum_{j=0}^{n} AR(j) \times w_{i,j} \tag{3.17}$$

Fiala et al. (2008) propose however to apply PageRank algorithm on author citation network. The frequency of co-authorship $\sigma_{v,u}$ is used instead for citation

[7] http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html

weighting in order reduce the importance of self citation and inter-collaborator citations. The impact of authors is defined by the following equation.

$$R(i) = \frac{1-d}{A} + d \sum_{(v,u)\in E} R(v) \frac{\sigma_{v,u}}{\sum\limits_{(v,k)\in E} \sigma_{v,k}} \tag{3.18}$$

In the same way, Deng et al. (2008) propose to compute the weighted PageRank algorithm on author co-citation network. This algorithm identifies important authors being co-cited with important authors too.

3.3.3 Scientific impact in the social web

The previously discussed metrics use mainly citation data to evaluate the scientific impact of publications and researchers. With the emergence of digital libraries and scholarly social networking services on the Web, new sources of information could be used to compute the impact of journals and authors. Some digital libraries such as ACM DL[8] reports the number of downloads in the last week, month and all the time. Some Web tools such as CitedIn[9] and ImpactStory[10] presents a set of featured social impact metrics for scientific purpose based on social bookmarks, blog comment, and citations from Wikipedia. Priem et al. have studied the validity of this data typically as up-to-date metrics for evaluating science (Priem and Hemminger, 2010; Priem and Costello, 2010; Priem et al., 2012a,b).

In practice, Haustein and Siebenlist (2011) proposed a set of impact metrics based on bookmarking data. For instance, *Usage Ratio* computes the ratio of bookmarked articles of a particular journal. *Usage Diffusion* corresponds to the number of users who bookmarked one of the journal's articles. Finally, *Article Usage Intensity* and *Journal Usage Intensity* are defined as the average of user's bookmarks for a particular article or journal.

Based on microblogging data, Eysenbach (2011) proposes a set of metrics for social impact and knowledge translation. In particular, *Twimpact Factor* computes the mean number of blogs that cite a particular article with a defined period. *Tweeted Half-Life* measures the necessary time before half cumulative number of blogs is published. Finally, *Twindex* computes a ranking score for each article in respect to *Twimpact Factor*.

[8] http://dl.acm.org/
[9] http://citedin.org/
[10] http://impactstory.org

3.4 Retrieval and ranking in digital libraries

The concept of digital libraries was introduced in 1945 by Bush and Think. The aim of such system is to provide efficient and flexible access to stored collection of books, records and communications. Nowadays, digital Libraries are attracting a heterogeneous population of information seekers in terms of age, knowledge and expertise (Borgman, 1996). In order to handle this issue, a variety of features have been incorporated into digital libraries design namely into information retrieving and ranking processes. We discus in this section the main retrieval approaches for digital libraries.

3.4.1 Document focused retrieval

Scientific publications and reviews were used as reference collections in first information retrieval systems (Salton, 1991). Retrieval models that were tested primarily on scientific collections have been integrated within digital libraries. These models assimilate relevance to the textual similarly between document and query. However, relevance in literature retrieval context does not depend only on topical similarity and may involve other criterions related to documents.

Giles et al. launched in 1998, the first search engine and digital library for academic research known as CiteSeer. This system ranks scientific papers by their hub score (Kleinberg, 1999) as good hubs may represent *"good introductions to areas of the literature in their prior work"* (Lawrence et al., 1999a). In the same way, scientific impact metrics measuring document quality (see section 3.3.1) can be used to rank retrieved documents. For instance, Sun and Giles (2007) propose to rank query results by their citation count, PageRank score and authority score of HITS as well as other metrics computed on the citation network. These approaches focus on the document level and integrate both content and referral dimensions.

3.4.2 Expertise oriented search

Finding experts in a particular domain is a common issue for scientific research. It helps to target potential collaborations, to identify keynote speakers, to select peer-reviewers and, in particular, to search relevant and quality papers. While browsing the search results, researchers make sure to include in their reading a list of credible papers by major authors and experts in their research field. In keeping with this search behavior, literature retrieval in digital libraries is viewed as expert search problem.

Expert search is in fact an information retrieval task that aims to rank experts in a given topic, rather than documents (Soboroff and Craswell, 2007). This task has received an increasing interest in both academic and industry (Deng

et al., 2012). Works proposed in this area are classified into two main categories (Serdyukov and Hiemstra, 2008):

Profile-centric approaches (Liu et al., 2005b; Petkova and Croft, 2006) models candidate expert by merging all corresponding documents into one personal profile. Standard information retrieval measures are then used to rank profiles in regard to the query. Let S_E be the profile set of expert E. A document D is included in S_E only if it mentions respective expert E. The relevance of an E in respect of a query Q is defined as follows (Petkova and Croft, 2006):

$$P(Q|E) = \prod_{i=1}^{|Q|} \sum_{D} P(q_i|D)P(D|E) \tag{3.19}$$

Document-centric approaches (Balog et al., 2006; Macdonald and Ounis, 2006; Fang and Zhai, 2007) propose to compute summarized score for each candidate expert based on related documents. The probability that an author a_i is expert with respect to query q topic is given by next formula (Balog et al., 2006).

$$p(a_i|q) = \sum_{d \in \mathcal{D}_{a_i}} p(a_i|d)p(d|q) \propto \sum_{d \in \mathcal{D}_{a_i}} p(a_i|d)p(q|d)p(d) \tag{3.20}$$

Where $\mathcal{D}_{a_i}$ is the subset of documents authored by candidate expert a_i.

3.4.3 Leveraging author social network

Recently, the problem of ranking literature has been dealt from a social point of view where social information networks are used to represent bibliographic resources. In contrast of document and expertise modeling, social approaches take into account the social relationships that involve authors and documents. The motivation behind the use of social links is to evaluate the relevance of bibliographic resources in their social context.

Social enhanced search for digital libraries investigates the social position of authors in order to rank bibliographic entities. As a matter of fact, relevant resources may be related to relevant persons in the social network. With respect to the social information network structures, proposed models in this context either compute a social relevance score of each author then transmit it to related documents or rank authors and documents jointly in a heterogeneous network. The two categories of approaches are discussed next.

3.4.3.1 Social relevance: from author to document

As discussed in section 3.4.2, the relevance of an author has been interpreted by his expertise in the related research field. However, author relevance in a broader

context may include several properties. In fact, a relevant author may show some authority in the social network but also characterized by its proximity to other social network nodes. In the same way, influencer and innovative researchers may receive as much attention as popular experts and domain pioneers. While text-based approaches still limited to identify these key actors, social network analysis provide a promising framework to distinguish relevant authors with a particular position in the social network.

The social relevance of authors in bibliographic networks is evaluated using network analysis measures introduced by both domains of social network analysis (Wasserman and Faust, 1994) and hyperlink analysis (Brin and Page, 1998; Kleinberg, 1999). Applied on author social network, Betweeness centrality identifies interdisciplinary authors who connect different sub-communities. Closeness centrality reflects the reachability and the independence of an author from his social neighborhood. PageRank algorithm and authority score of HITS algorithm allow to distinguish authoritative authors in the social network. Finally, Hub score of HITS algorithm identifies central authors with an important social activity involving authoritative colleagues.

Beside these social centrality measures, social-based metrics of authors' scientific impact introduced in section 3.3.2 may be used in this context to evaluate the importance of authors, namely weighted PageRank (Fiala et al., 2008), co-citation PageRank (Ding et al., 2009) and AuthorRank (Liu et al., 2005a). Other measures that addressed advanced properties of social network actors in a boarder context could be applied to identify important author in the social network such as network experts (Li et al., 2007; Karimzadehgan et al., 2009), influencers (Tang et al., 2009; Weng et al., 2010) and actors' similarity (Jeh and Widom, 2002; Gollapalli et al., 2012; Chen et al., 2011b).

The interest of identifying key authors in the social network is to rank retrieval results according to the relevance of related authors. Empirical studies (Kamps, 2011) confirm the effectiveness of such approach. However, propagating social relevance scores from authors to documents sill a common issue of these approaches specifically in the case where the document is written by several co-authors. Obviously, standard aggregation operators can be used to estimate the social relevance of documents, for instance, the mean of authors scores (Kirchhoff et al., 2008).

$$r_d = \frac{1}{n_d} \sum_{a_i \in \mathcal{A}_d} \mathcal{C}_{a_i} \tag{3.21}$$

Where $\mathcal{A}_d$ represent the set of related authors. n_d is the co-author number. $\mathcal{C}_{a_i}$ is the social relevance score of author a_i.

We note that the score aggregation function may impact the retrieval performances as it should give different interpretations of social relevance. In the case of literature search, the sum of the authors score may show better results

(Kamps, 2011). This is explained by the fact that both number of authors and their high relevance are indicators of document quality.

Unfortunately, ranking documents using only the social score of authors may lead to a topic drift over the search results. In other words, documents of highly relevant authors are always ranked among top results even though they do not discuss the query topic. To tackle this problem the social score of documents r_d and the topical relevance of documents with regard to the query $rel(q, d)$ are combined using the next formula (Kirsch et al., 2006).

$$rel_f(q, d) = r_d \times rel(q, d) \tag{3.22}$$

In the same context, the social relevance of authors may deserve less or more importance in comparison to topical relevance. A normalizing dumping factor λ can be used in this case in order to tune the impact of the social relevance on the ranking process (Kirchhoff et al., 2008).

$$rel_f(q, d) = \lambda\ r_d + rel(q, d) \tag{3.23}$$

While linear combination is confronted to the problem of relevance distribution of query independent features (Craswell et al., 2005), this problem can be solved by adjusting relevance scores using appropriate transforming functions (*e.g.*, log scale, sigmoid). Learning to rank models (Liu, 2009) can be used instead to combine topical and social relevance. However, a training process needs to be conducted first. Another alternative is to model authors and documents with heterogeneous networks then rank jointly the two entities. This approach is discussed in the next.

3.4.3.2 Co-ranking documents and authors

Previously cited approaches estimate independently the social relevance of authors then propagate it to related documents. However, authors and documents are closely related entities that mutuality reinforce the relevance of each other. Based on this idea, recent works propose to jointly rank documents and authors in heterogeneous bibliographic networks (Zhou et al., 2007; Sayyadi and Getoor, 2009; Yan et al., 2011).

Zhou et al. (2007) propose a ranking algorithm that identifies relevant documents written by reputable authors and vice versa. In particular, a PageRank-like algorithm with two personalized random walks is applied on the bipartite graph of authors and documents. Intra-class random walk estimates the local authority of authors and documents. Intra-class randomwalk models the mutual reinforcing between documents and authors.

Sayyadi and Getoor (2009) propose a ranking algorithm, called FutureRank, that predicts future PageRank scores for authors and articles. This algorithm

applies in the first step a PageRank algorithm on the citation network and HITS algorithm on the bipartite authorship network then combines the two results.

Yan et al. (2011) propose to rank articles, authors and journals in heterogeneous networks. The defined algorithm, P-Rank, estimates the prestige of an article based on the prestige of citing articles, authors and journals. On the other hand, the prestige of authors and journals is estimated based on the prestige of related papers. Similarly to the co-ranking algorithm proposed by (Zhou et al., 2007), P-Rank uses two random work probabilities. First, intra-class walk is applied on citation links between articles. Second, inter-class walk models the probability to move from a particular article to the related author or journal.

3.4.4 Social and collaborative search

Digital libraries usually provide professional metadata about articles such as general subjects and keywords. However, this metadata is not more effective for literature search than user-generated content provided by social networking services (Koolen et al., 2012). The effectiveness of social generated data is confirmed by empirical studies (Heymann et al., 2008b,a; Yi and Chan, 2009) that have shown its promising role as potential information retrieval tools, and for the design of such systems. In fact, tags and other social-generated data such as comments, reviews, rates, social networks and query logs provide as an implicit feedback about retrieval relevance which is not always available in the content.

Some few works have addressed the problem of social and collaborative retrieval over digital libraries. Koolen et al. (2012) suggest to use tags and user reviews from forums in order to improve book search. They propose to index user-generated data as well as regular content and other professional data using traditional information retrieval techniques. Experimental results show high retrieval performances by forum reviews. A similar approach is proposed by Jomsri et al. (2009) where tags are used to index scientific papers. Sun et al. (2008) propose to extract implicit feedback from user query logs for personalized ranking of search results.

Beside these works that propose a practical way to conduct social and collaborative search, Evans and Chi (2008) define an innovative search model that takes into account all implicit and shared information provided by users before, during, after search. However, such models are not yet explored by current research.

3.5 Academic search engines: social features in focus

Academic search engines and digital libraries enable researches to access to a large collection of multidisciplinary publications. Regardless of the amount of covered publications, these tools provide a variety of innovative retrieval components and socially-aware interfaces. We present in this section a comparison of the main academic search engines and digital libraries particularly in terms of integrated social features and information access and retrieval.

Academic search engines include a wide range of applications: (*i*) Web search engine such as Google Scholar[11], Microsoft Academic Search[12] and CiteSeerX[13]; (*ii*) bibliographic databases such as DBLP[14]; (*iii*) search and analysis tools such as AMiner[15], formly ArnetMiner; (*iv*) and scholarly social bookmarking services such as CiteULike[16], Mendeley[17] and BibSonomy[18]. A comparison of these tools is presented in table 3.1.

At the article level, the three academic search engines provide information about citation data, namely the list of referenced articles as well as set of citation metrics. Digital libraries such as Mendeley and AMiner provide however a partial access to this data. All the compared tools, expect DBLP, are able to recommend related publications to a particular article.

The majority of academic search engines enable author-based navigation of articles. One of the basic features is to list publication by author name. Scholar, Academic Search, CiteSeerX and AMinder build a structured and a rich profile of authors including affiliation, research topics and impact metrics. While co-authorship data is widely supported, author citation networks is only supported by Microsoft's Academic Search. This tool provides an innovative way for citation network exploration and interrogation.

Scholarly social bookmarking services is distinguished by promoting user based navigation and search of literature. User tags are used in this context as a key feature for article classification. In addition, scholarly social bookmarking services build detailed profiles for users including personal information and tagging activity. These tools provide also a personalized access to literature through a friendship network. We note that user profile and friendship networks are supported by some academic search engines such as Google Scholar though connected services or authenticated authors.

[11] http://scholar.google.com/
[12] http://academic.research.microsoft.com/
[13] http://citeseerx.ist.psu.edu/
[14] http://www.informatik.uni-trier.de/
[15] http://arnetminer.org/
[16] http://www.citeulike.org/
[17] http://www.mendeley.com/
[18] http://www.bibsonomy.org/

		Scholar	Academic Search	CiteSeerX	DBLP	AMiner	CiteULike	Mendeley	BibSonomy
Article	Citation Count	■	■	■	□	■	□	■	□
	Citation Network	■	■	■	□	□	□	■	□
	Related Publications	■	■	■	□	■	■	■	■
Authors	Profile	■	■	■	□	■	□	□	□
	Publications	■	■	■	■	■	■	□	■
	Co-authors	■	■	■	■	■	□	□	□
	Citations	□	■	□	□	□	□	□	□
Users	Profile	■	□	□	□	■	■	■	■
	Tags	□	□	■	□	□	■	■	■
	Friendship	■	□	□	□	■	■	■	■
Access & Retrieval	Time and trends	■	■	■	□	□	□	□	□
	Keywords and tags	□	■	□	□	□	□	□	■
	Venues	□	■	□	■	■	□	□	□
	Organisations	□	■	□	□	□	□	□	□
	Expertise	□	■	■	■	■	■	□	□
	Similar users	□	□	□	□	□	■	■	□

Table 3.1: Comparison of academic search engines and digital libraries

Academic search engines and digital libraries present a variety of approaches for literature access and retrieval. In this context, expert search represents the most common feature by the compared tools. Academic Search stands a first search tool in terms of supported features including temporal analysis, trend detected, keywords and tag search, venue and organization classification and expertise search. User search and recommendation is however presented by some few social bookmarking services, namely CiteUlike and Mendeley.

Conclusion

We presented in this chapter an overview of social network approaches for literature access. In particular, we have discussed main social network models that represent authors of scientific publications as well as annotators of scholarly social bookmarking networks. Furthermore, we presented principal metrics of authors' scientific impact, mainly social network based metrics. Finally, we discussed social network approaches for retrieving and ranking literature.

One common feature of the discussed approaches in this chapter is that they exploit the social network structure in order to enhance the retrieval process over traditional documents. In the next chapter, we will focus on social retrieval approaches within social networking data, namely microblogs. This data presents different formats and properties in spite of bibliographic resources.

Chapter 4

Information retrieval in microblogging networks

Introduction

User generated content, such as microblogs, are more and more available on the Web. Users are aware of the availability of this data. They express an information need to access to this information. For instance, users are interested in new updates about an event as wall as people opinion about. This type of information is not available via traditional Web often providing professional information regardless of current events and the social interest.

Microblogs, a short form of blogs, stand as promising tool where users can find recent information published by other users. These social networking websites are distinguished by information diversity as well as the intensive social interaction within. People can in fact tag their blogs, post comments, and share further resources such as photos, videos and Web pages. This information is usually not yet indexed and thus not available via classical search engines.

Searching information within microblogs differs from Web search since the searched data differs in content and format as well as the underlying motivation. In contrast of Web search, where queries are submitted for informational, transactional or a navigational propose, search within microblogging social networks, is motivated by the social activity of the person as well as current event and trends that inspire microblogging community. Relevance in this context dependent on the microblogging intention and includes several relevance factors typically the social context of the information.

In fact, people in microblogs have the same importance than information. The quality of information is defined by the quality of authors and vice versa. In this

context, key microbloggers in the social network are identified in order to access relevant information. These key microbloggers correspond mainly to network influencers that influence microblogging activity.

In this chapter we discuss social network based approaches for microblog retrieval. First, a brief introduction to microblogs is presented. After that, we discuss main information retrieval tasks in microblogs then we later focus on social approaches for microblog search. Furthermore, we discuss main approaches for identifying network influencers then we conclude with a brief description of TREC microblog evaluation campaign.

4.1 Overview of microblogs

A microblogging service is a communication medium and a collaboration system that allows broadcasting short messages. In contrast to traditional blogs, media-sharing and social networks services, microblogs are textual messages submitted in real-time to report an actual interest. Often limited to few characters, it is practical to create a microblog from mobile devices to instantly report a real world event. Java et al. (2007) defined microblogs as a self motivated communication service.

> *These tools provide a light-weight, easy form of communication that enables users to broadcast and share information about their activities, opinions and status (Java et al., 2007).*

Besides this personal and entertaining purpose, microblogs have recently attracted the attention of corporations (Riemer and Richter, 2010; Zhang et al., 2010; Müller and Stocker, 2011), online communities and news editors as a promising tool for team collaboration and information broadcasting. Having these various usages, three main categories of microblogs are identified:

Information broadcasting: This category of microblogs includes mainly real-time news. The intention behind these messages is to largely spread an information though the social network. Broadcasted messages announce either a personal information such as *"The Smith family has a newborn!"* or report a large-scale news such as *"Gulf Oil spill continues to grow and spread east"*.

Communication: This category of microblogs communicates thoughts, emotional and factual information such as *"No matter what people say"*, *"I'm so sad!"* and *"The moon is the Earth's satellite"*. It includes also status updates from other social networking services reporting latest activities such as auto-generated YouTube activity status *"I liked a @YouTube video http://youtu.be/..."* .

Collaboration: This category includes community notifications, group discussions and questioning-answering posts such as *"A new release of Twitter API"*, *"Issue allocating memory"* and *"No, only Android devices are sup-*

ported". Collaborative microblogs include also posts sharing helpful web resources and exchanging knowledge such as technical tips. Target audience are usually mentioned in this category of microblogs.

Several microblogging services are available on the Web namely Dailybooth[1], FriendFeed[2], Tumblr[3]. These Web sites allow friends to share their interest. Twitter[4] stands actually as the most popular microblogging service with over 200 million active users and 400 million microblog per day[5]. In this thesis, we are particularly interested in Twitter microblogging service.

4.1.1 Twitter on focus

Launched in March 2006, Twitter provides innovative features in comparison to other social networking Web sites and microblogging services. A microblog, known as *"tweet"* in Twitter parlance, is a plain text blog with up to 140 characters. This format ensures microblogs compatibility with other services such as mobile applications and Short Message Service (SMS). The Twitter social network is based on the principle of followership. A microblogger may follow another one, known as *"following"*, and followed in his turn by someone else, known as *"follower"*. Once authenticated, tweets of followed people (*followings*) are displayed to the microblogger in a reverse chronological order. Microbloggers can access to tweets from other persons in the social network unless no restriction is applied to their tweets. By default, tweets are publicly visible.

Microbloggers represent individual persons but also institutions, companies, communities and online services. Each Microblogger is identified with a *@username*. Figure 4.1 shows some popular tweets from United States presidential election in 2012.

Each tweet is associated to one microblogger who actually represent the author of the tweet. Meanwhile, a tweet can be reblogged by another person for instance, as illustrated in figure 4.1, tweet *(e)* originally authored by *Barack Obama* and reblogged by Twitter Chairman *Jack Dorsey*. Reblogging mechanism is known in Twitter as *"retweeting"* and it is assimilated to *"sharing"* concept in other social networking services. Tweet *(a)* by *Barack Obama* represents in fact the most reblogged tweet in Twitter with over $800,000$ retweets. This tweet is also saved $301,873$ times as *"favorite"*. This feature is similar to *"like"* in other social networking services.

Twitter enables users to attach photos to their tweets as shown in tweet *(a)*. In practice, images, videos and web pages can be attached by adding corresponding URL to the tweet text. Similarly to social bookmarking service, Twitter enables

[1] http://dailybooth.com/
[2] http://friendfeed.com/
[3] http://www.tumblr.com/
[4] http://www.twitter.com
[5] http://blog.twitter.com/2013/03/celebrating-twitter7.html

Figure 4.1: Popular tweets from United States presidential election, 2012

users to annotate their own tweets with personalized tags known as *"hashtags"*. A hashtag is a non-spacing word with prefix character *"#"*. For example, tweet *(b)* is tagged with hashtag *"#election2012"* which is dedicated for election-related news.

Tweets *(c)* and *(d)* give two examples of conversations mechanism in Twitter. These tweets are published in interaction with *Barack Obama* tweet *(a)*. A tweet will be addressed to a particular user if his *@username* is mentioned at the beginning. Such tweet is called *"reply"* if only it points to a previous tweet as the case of tweets *(c)*. In the second tweet *(d)*, British Prime Minister *David Cameron* congratulates *Barack Obama* with a new tweet. *Barack Obama* is mentioned in this tweet. Being mentioned in the tweet, a microblogger should be concerned about. Unlike tweet reply, no discussion board is built for mentioning tweets unless it answers another tweet.

4.1.2 Characterizing microblogs

In order to understand microblogging activities, several works have investigated linguistic, topical, spatiotemporal, demographical, and topological properties of the Twitter social network. We present in this section main findings and we

discuss major challenges to microblog retrieval.

Linguistic analysis on several tweet datasets with different sizes (Jansen et al., 2009) shows that tweets' length in terms of words is almost similar. The average number of words per tweet is taken between 14 and 16 as presented in table 4.1. In comparison to Wikipedia articles where average length is about 320 words per article, tweets are extremely short. The Tweet length is actually comparable to sentence level where median length is above 15 words (Kornai, 2008). This property makes information retrieval and access over microblogs a challenging task often similar to sentence retrieval rather than document retrieval. Basic retrieval concepts such frequency-based weighting are no longer appropriate with respect to fine retrieval granularity.

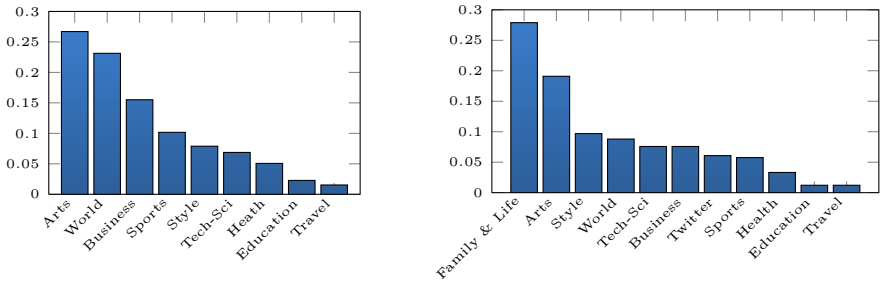
	Tweet length (words)			Tweet length (characters)		
Tweet measures	14.2 K tweets	38.8 K tweets	2.6 K tweets	14.2 K tweets	38.8 K tweets	2.6 K tweets
Average	15.4	14.3	15.8	86.3	89.1	102.6
SD	6.8	6.4	6.6	36.5	35.3	36.4
Max	33	33	43	142	155	185
Min	1	1	3	1	1	16

Table 4.1: Linguistic statistics for tweets (Jansen et al., 2009)

Analyzing a dataset of 1.2 billion tweets and 11,924 articles from New York Times, Zhao et al. (2011) report that Twitter have a similar topic distribution to traditional media as shown in figure 4.2. Nevertheless, *Family & Life* category, which does not appear in New York Times articles, represents the dominated category in Twitter. This category of tweets is related to family, daily activities, emotional status, *etc.* *Arts* and Style categories, typically celebrity related tweets, are also strongly present in Twitter. Accordingly, a new user's information need that tends towards self and socially motivated purpose is expected in microblogs. On the other hand, some studies[6] report that, despite the low spam ratio in Twitter (3.7%) as result of the efficient spam policy, the majority of tweets are in fact "pointless bubble" (40.55%) or conversational (37.55%). This undesirable content may present a challenging issue for microblog retrieval systems.

In order study the information diffusion in Twitter, Kwak et al. (2010) have conducted a large scale analysis of Twitter social network including 41,7 million users. As shown in figure 4.3(a), user activity increases according to the number of his followers. A low activity is noted for users with less than 10 followers. Conversely, higher number of tweets is expected for highly followed microblogger. Furthermore, the number of reciprocal followerships is positively

[6]http://www.pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf

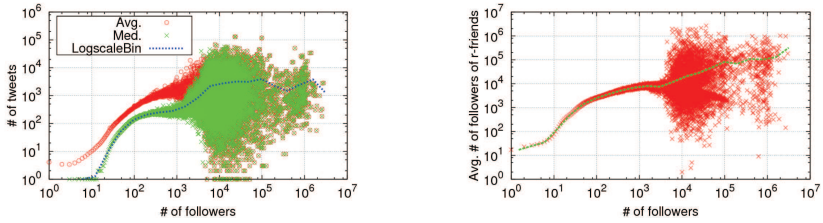


(a) New York Times

(b) Twitter

Figure 4.2: Distribution of topic categories in New York Times and Twitters (Zhao et al., 2011)

correlated with the number of followers typically for users with less than $1,000$ followers as shown in figure 4.3(b). A dispersed number of reciprocal followerships is noted above this number. This is explained by the fact that some users adopt an excessive behavior to increase their popularity. Thus, microblogging retrieval is confronted to another issue that consists on information credibility and user authority in the social network.



(a) The number of followings and that of tweets per user

(b) The average number of followers of r-friends per user

Figure 4.3: Network properties of Twitter Social Network (Kwak et al., 2010)

Analyzing microblog distribution over, tweets seems to have similar distribution across week days as show in figure 4.4 (McCreadie et al., 2012). An important activity is however noted in the second half of the day. Furthermore, activity peaks are registered at some particular days which correspond in fact to large scale events. In the aim of studying trend patterns, Benevenuto et al. (2010) have compared the daily frequency of tweets related to music artist *Michael Jackson* and those tagged with *#musicmonday*. As shown in figure 4.5(a), tweets distribution presents a peak on June, 25^{th} which corresponds to the celebrity's death. A second peak is registered few days later when media reported new details about the health status of Michael Jackson. 20 days later, an important decrease of tweets activity marks the end of this trends. While some trends

disappear few days after their birth, others remain active for a long time. As shown in figure 4.5(b), *#musicmonday* trend follows however different pattern with weakly recorded peaks. One challenge for microblog retrieval is to detect rising trends and predicts their life time in order to present, at appropriate time, interesting content for users.

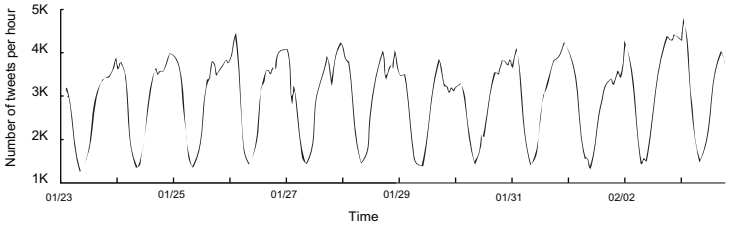
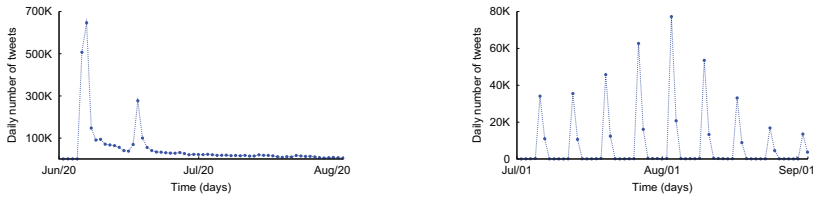


Figure 4.4: Twitter distribution over time (McCreadie et al., 2012)



(a) Michael-jackson

(b) #musicmonday

Figure 4.5: Daily number of tweets of analyzed events (Benevenuto et al., 2010)

4.2 Retrieval tasks in microblogs

Microblogging data is more and more available on the Web. This huge amount of social stream has created new information retrieval tasks that correspond to new user's needs of information. We identify in the next the main retrieval tasks over microblogs.

4.2.1 Real-time search

Seeking for information over microblogs helps to find reliable, concise and real-time information about a recently happened event (few seconds ago up to few days) (Mills et al., 2009). It would take a time before this information became available on the Web and then indexed by search engines (Dong et al., 2010). Microblog real-time search task is an ad-hoc retrieval task where users are interested in most recent and relevant information (Ounis et al., 2011). Formally,

real-time search task is defined by the probability $P(t|q, \theta_q)$ where t refers to the tweet, q is the user query and θ_q is the query time. Microblog, as well as blogs in a broader context, are treated here as independent documents and they are ranked in reverse chronological order along with their relevance to the query (Mishne, 2006; Thelwall and Hasler, 2007). Empirical studies for microblog search show relevance of tweet may depends on several features in addition to the textual similarity to the query such as the number of followers and followings, the freshness of information, included URLs and the user's location (Nagmoti et al., 2010; Duan et al., 2010; Sankaranarayanan et al., 2009). These features are discussed in more detail in the next section 4.3.

4.2.2 Followings suggestion

Followings suggestion enhances the microblogging experience by recommending like minded people. Hence, microbloggers can easily find interesting people who share the same interest. In fact, followings suggestion is a recommendation task that selects for a particular user a set of users to be followed. Followings recommendation is based either on the user's profile or on the social network topology. In the first category, recommender systems match user information and tweet history to suggest similar users discussing related topics. Armentano et al. (2011) propose to conduct recommendation based on content-based similarity as defined next:

$$sim(u_C, u_T) = \max_{\forall i: f_i \in followees(u_T)} sim_{cos}\left[profile_{base}(f_i), profile_{base}(u_c)\right] \quad (4.1)$$

with sim_{cos} is the cosine similarity between two the microbloggers u_c and f_i.

In the second category, recommender systems suggest new followings from the social neighborhood or people followed by similar users. Accordingly, Armentano et al. (2012) propose to count the number of friends between the two microbloggers as defined next:

$$w_c(x) = |i \in followers(x) \cap i \in followers(U)| \quad (4.2)$$

Followings suggestion is primarily conducted in regard to a specific microblogger. This task can be driven with respect of a particular topic. In this case, recommendation is viewed as a retrieval task where users search for relevant people discussing a particular topic. In this context, Hannon et al. (2010) propose a query-based retrieval and profile-based recommendation system that assigned to the microblogger's profile a set of representative terms as defined in the next equation:

$$TF - IDF(t_i, U_T, U) = tf(t_i, U_T) \times idf(t_i, U) \quad (4.3)$$

where $tf(t_i, U_T)$ is the frequency of term t_i in user profile U_T and $idf(t_i, U)$ is the inverse document frequency of t_i in the rest of profiles U.

Regardless of recommendation method, followings suggestion may verify also some key properties of microbloggers such as credibility, authority and expertise (Ting et al., 2012; Garcia and Amatriain, 2010).

4.2.3 Trend detection and tracking

Trends detection aims to identify active topics in the social network (Mathioudakis and Koudas, 2010; Becker et al., 2011). In particular, trend detection task search for frequent expressions in microblog stream and infer public interest within a particular period. Trends are mainly generated from the last few days of microblog stream.

Being correlated with real world events, a trend correspond, in a border context, to a large scale events that interest a wide range of users such as political events (Tumasjan et al., 2010, 2011) and sport event (Nichols et al., 2012; Lanagan and Smeaton, 2011). The amount of published microblogs helps to quantify the audience and qualify the people reaction. Monitor a specific event over microblog stream may provide early warnings in emergency situations such as earthquakes (Sakaki et al., 2010; Earle et al., 2012, 2010) and epidemics (Lampos et al., 2010; Aramaki et al., 2011). In this context, Sakaki et al. (2010) propose first to tack tweets using a set of keywords then apply Kalman filtering to estimate the location of related event.

4.2.4 Opinion and sentiment retrieval

Opinion retrieval aims to extract and rank opinioned tweets (Zhang et al., 2007). Likewise opinion retrieval over regular blogs, relevant tweets must satisfy the user's information needs and express at the same time a clear opinion about the query topic regardless of its polarity: negative, positive or mixed. Opinion expressed in the tweet is static. However, the position of related microblogger changes over the time under the influence of his social neighborhood (Boyd et al., 2010; Romero et al., 2011).

Although some works have addressed opinion retrieval (Luo et al., 2012), sentiment analysis is more relevant in microblog context (Go et al., 2009; Bollen et al., 2011; Kouloumpis et al., 2011; Barbosa and Feng, 2010; Jansen et al., 2009; Agarwal et al., 2011). In fact, tweets are too short to express an opinion but can communicate however much sentiments. A wide range of work has interested in this problem typically for identifying people sentiment about a movie, a brand, a political candidate, etc. Go et al. (2009) propose to use machine learning classifiers to detect tweet sentiment where each term is considered as a feature. Results show that Support Vector Machines (VSM) presents higher results with an an accuracy of 82.9%.

Sentiments are commonly balanced between negative and positive but it can

include more emotional states such as surprise, fear, disgust and anger (Roberts et al., 2012).

We focus in two following sections 4.3 and 4.4 respectively on microblog search and the identification of microblog influencers since our contributions mainly address these two issues. Microblog search is in fact a generalized task of real-time search task. Indentfying microblog influencer is viewed as a sub-task of followings suggestion task.

4.3 Microblog search

Microblog search is an emerging research topic that has acquired more attention with the increasing popularity of microblogs. Efron (2011) distinguished two types of microblog search namely *"asking"* and *"retrieving"*. *Asking* for Information is similar to questioning and answering activities (Q&A) where people post questions via microblogs and seeks for community help. Microblog *retrieving* is indeed compared to traditional ad-hoc retrieval task were people seek for relevant microblogs with regards to their search queries.

> *Retrieving is similar to traditional, ad hoc IR. Interactions of this type are likely to involve a "query" that is posed against an index of microblog data (Efron, 2011).*

This section focuses on a microblog retrieving task that we refer to as *"microblog search"*. Despite real-time search that we discussed in section 4.2.1, microblog search is a generalized ad-hoc retrieval task not restricted to a time frame or a presentation format. One microblog search systems may rank tweets regardless of their freshness as required by real-time search. Other systems may display results as a word cloud instead of tweet list (Efron, 2011).

4.3.1 Search motivations

The main motivation of an information retrieval system is to satisfy the user's information needs. Before defining what are the relevance features that characterize microblogs, it is primordial to determine what are the factors that motivate user search within microblogs. Related to this context, Broder (2002) and Manning et al. (2008) identified three types of search queries that reflect search motivations on the web namely, navigational, informational and transactional queries. Comparing web search to microblog search, Teevan et al. (2011) identified three main factors that motivate microblog search including socially search, temporal search and topical search. The three search motivations are detailed next.

Social information. An important part of microblogs search (26%) is motivated by a social intent. This is realized through search queries mentioning the name of a person or through the social interaction that have triggered search session (Evans and Chi, 2008). In particular, users of microblogs are interested to search for like minded people or to discover what a person is talking about. In addition, public trends may encourage people to launch a search query and learn more about people opinions, sentiments and reactions.

Temporal information. As microblog have essentially a temporal intent particularly to get informed about the last news and events, typically if this information is not yet available on other web resources, search over microblogs have particularly a temporal motivation. Users are searching for information related to news and trending topics (Phelan et al., 2009; Dong et al., 2010; Mathioudakis and Koudas, 2010). Temporal search helps also to get a real-time summary from the event's site (Hughes and Palen, 2009; Diakopoulos and Shamma, 2010). Furthermore, people may use microblogs to search for last updates about regional and local information such as weather forecast, traffic jam, police instructions or service status (Yardi and Boyd, 2010; Sankaranarayanan et al., 2009).

Topical information. Similarly to Web search, microblog search shows a topical interest (Sousa et al., 2010) with more or less focused queries such as *"astronomy"* or *"black hole"*. Despite the temporal or the social intent, topical search helps to find an information about an old trends and events or to discover users with a particular interest outside the followings network.

4.3.2 Relevance factors

Microblog relevance is a composite concept that includes several factors. Some work have studied relevance dimensions in a boarder context (Cosijn and Ingwersen, 2000). Among the discussed features, contextual relevance is the mostly used in microblogs context. Contextual relevance includes in fact different types of features. We identify in figure 4.6 the main relevance features for microblog search. These features are discussed next.

Content. This relevance feature stands as the primordial feature for microblog search, typically for ad-hoc search. The frequency of terms is used in this context to estimate the similarity between the tweet and the query. Traditional information retrieval model such as the Boolean Model, the Vector space model and the probabilistic model are also applicable. Ferguson et al. (2012) have studied the impact of frequency normalization and tweet length through BM25 ranking model. Results show that the outcome of term frequency weighting is minor while document length has a negative impact on microblog retrieval

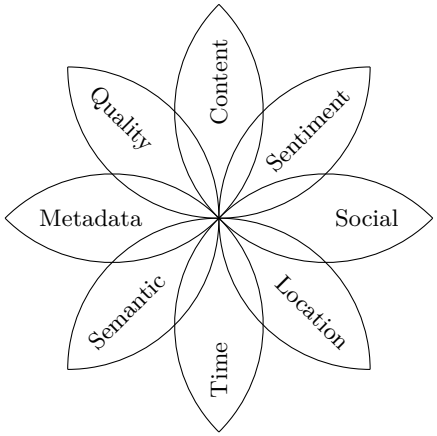


Figure 4.6: Microblog relevance factors

effectiveness. Advanced techniques are used in this context to overcome this problem namely term co-occurrences (Lin et al., 2012), text classification (Sriram et al., 2010) and pseudo-relevance feedback (Massoudi et al., 2011; Bandyopadhyay et al., 2012; Zhang et al., 2012). Zhang et al. (2012) have compared the effectiveness of a set of well known baselines for microblog query expansion. Results show that Kullback-Leibler divergence model present better results in comparison to BM25, PL2 and Kullback-Leibler divergence language model with Dirichlet smoothing.

Quality. Pointless, ambiguous and spam tweets should not affect microblog search (Ross et al., 2010; O'Connor et al., 2010; Chu et al., 2012). Several works have addressed tweet quality issue at pre-retrieval and post-retrieval process. In order to keep only quality tweets, several features are taken into consideration such as the tweet length (Jansen et al., 2009), the language (Bergsma et al., 2012), the out of words vocabulary (Han and Baldwin, 2011; Gouws et al., 2011; Duan et al., 2010), the emoticons (Davies and Ghahramani, 2013; Cui et al., 2011) and the punctuations (Green and Sheppard, 2013), etc.

Metadata. Microblog metadata includes hashtags, mentions and URLs. This data describe tweet content and context. Although some works consider the presence of user generated as a relevance indicator (Li et al., 2011a; Tao et al., 2012; Duan et al., 2010), other propose to exploit this information for query and tweet expansion typically using hashtags (Efron, 2010) and the content of attached URLs (Jin et al., 2011; McCreadie and Macdonald, 2013),

Semantic. Tweets that contain named entities related to the query such as a person, a place or a product would be more relevant for microblog retrieval

(Spina et al., 2012; Jung, 2012; Genc et al., 2011). For instance, tweet mentioning named entity *"Barack Obama"*, who is in fact one of the main candidates of the US presidential election 2012, are presumed relevant for this event. Recognized entities in tweet help to extends it with further information for instance, from Wikipedia. In order to identify named entities in tweets, Meij et al. (2012) propose to use n-gram features (*e.g.*, idf of the n-gramme in Wikipedia titles), concept features (*e.g.*, wikipedia category hierarchy) and tweet features (*e.g.*, hashtags). An evaluation track is held by INEX 2013 workshop in this purpose called *"Tweet Contextualization Track"*. Participating systems must provide further information about the subject of the tweet using Wikipedia documents. The goal of this task is to answer the question *"What is this tweet about?"* (Bellot et al., 2012).

Time. As discussed previously, microblog search is motivated by temporal information need. In this context, recently published tweets may be relevant. Accordingly, Metzler and Cai (2011) propose to consider time difference between the query Q and the tweet D as a learning to rank feature for tweet search system. Furthermore, tweets published on activity periods of the query topic, which usually correlate with the related event and trend, would be relevant than tweets published at anytime else. Based on this idea, several work propose to study the distribution of microblogs over the time and select relevant tweets from convenient periods that better reflect the query temporal interest (Choi and Croft, 2012; Kumar and Carterette, 2013; Miyanishi et al., 2013). For instance, Choi and Croft (2012) propose to compute the relevance probability of the time period t as the proportion of tweets of period t that were returned by the query q, noted $\#docs(t, D, Q)$, over the number of tweets published in this period, noted $\#docs(t', D, Q)$. The probability $P(t|D, Q)$ is defined by the next equation:

$$P(t|D, Q) = \frac{\#docs(t, D, Q)}{\sum_{\forall t'} \#docs(t', D, Q)} \tag{4.4}$$

Location. The location where the tweet is published helps to estimate its relevance. For instance, a tweet coming from the place where the ceremony of the Olympic Games is hold seems to be more relevant as people may report fresh news and live photos (Yardi and Boyd, 2010). In addition, users searching for *"municipal elections"* may be interested of tweets from their region rather than tweets from other countries. In of view of that, location has been used by several microblog models in order to estimate tweet relevance as well as aggregating local and regional stream (Albakour et al., 2013; Lee et al., 2011). Besides, geographical metadata may be useful for topic modeling over microblogs. In this context, (Kotov et al., 2013) propose a geographically-aware extension of the *LDA* algorithm based on microblog geo-tags. Nevertheless, the used of geographical based features for microblog retrieval is limited by data sparsity. To

resolve this problem, (Kinsella et al., 2011) propose to exploit the language of the tweet and zip codes to detect the origin the tweet.

Social network. This category of features includes quantitative and qualitative metrics that describe the social significance of tweets and microbloggers. In fact, ranking tweets must consider the credibility of the information and present only reliable resources. This issue is addressed from users' point of view. The popularity and the trustworthiness of individuals are investigated in order to distinguish credible tweets (Ravikumar et al., 2012). Tweets are more likely to be relevant if they are retweeted by important users in the social network. In particular, Duan et al. (2010) propose to rank tweets according to the number of followers, the number of mentions as well as the authority of the microblogger, computed by applying *PageRank* algorithm on retweet social network. Furthermore, some works propose to use the social context to ensure a personalized ranking of search results (Uysal and Croft, 2011; Feng and Wang, 2013). In accordance to the topical interest of the microblogger, the number of followers and the retweeting and mentioning history, users who are likely to be retweeted are identified. Respective tweets are then ranked on the top of the result set.

Sentiment. Tweets that communicate sentiment about a person, a product or movie may be relevant in the case where users are interested on a summary of public sentiments. Some work proposes therefore to highlight sentiment tweets in the search results. In particular, Bermingham and Smeaton (2012) propose to filter search results by sentiments. Four ranking lists are provides: positive tweets only, negative tweets only, positive and negative tweets, and finally a random sampling. Results show that users of microblog retrieval systems are less interested in positive and mixed tweets. Conversely, negative and random tweets show better results. Other work considers that sentiment tweets are irrelevant by matter of subjectivity and propose to filter sentiment tweets typically that contain emoticons (Karimi et al., 2012).

Empirical studies on relevance factors show that the previous features have various impacts on tweet search effectiveness. Tao et al. (2012) show that semantic-based features namely the semantic overlap between the tweet and the query computed using *DBpedia* as well as the presence of URLs are highly effective for tweet search as presented in table 4.2. Conversely, replies and positive sentiment tweets show negative impact on the retrieval effeteness. Similar results presented by Damak et al. (2013) show the interest of keywords-based and URL-based features while hashtags, mentions and replies based features seem not effective for tweet search.

We note that a feature would be more or less effective according to the query topic, search motivation and retrieval task. For instance, location based feature

Feature Category	Feature	Coefficient
keyword-based	keyword-based	0.1701
semantic-based	semantic-based	0.1046
	isSemanticallyRelated	0.9177
syntactical	hasHashtag	0.0946
	hasURL	1.2431
	isReply	-0.5662
	length	0.0004
semantics	#entities	0.0339
	#entities(person)	-0.0725
	#entities(organization)	-0.0890
	#entities(location)	-0.0927
	#entities(artifact)	-0.3404
	#entities(species)	-0.5914
	diversity	0.2006
	sentiment	-0.5220
contextual	social context	-0.0042

Table 4.2: Relevant-tweet prediction model coefficient for employed features (Tao et al., 2012)

may be relevant to locally trending topic such as *"Carnival of Venice"* but not for topical query such as *"Ice Age"*. The last query that corresponds also to a movie name may have different interpretation if the movie *"Ice Age"* is actually playing in cinema. In order to overcome this issue, different relevance features may be combined to estimate a global relevance. This approaches are discussed in detail in section 4.3.4.

4.3.3 Indexing microblog stream

Indexing microblog data involve the same retrieval processes as traditional documents including tokenization, stopwords removal, stemming, etc. Previous works for these purposes is also applicable for microblog data. However, the amount of microblogs published every day (*i.e.*, over 400 million[7] tweets) has challenged the indexing and retrieving process. Although some retrieval systems may rely on search caches, query logs and personalized information retrieval to speed up search. Retrieval latency, data availability, index concurrency and temporal search remain critical for microblog search, typically for real-time search. Unfortunately, this issue has not been widely discussed in the literature.

The two main works in literature (Chen et al., 2011a; Busch et al., 2012) have addressed this issue from two points of views. Busch et al. (2012) have investigated the problem of indexing real-time stream from a software engineering perspective of view. They propose to split the index over several servers where only one instance is actively modified. The user query is analyzed by a front

[7]https://blog.twitter.com/2013/celebrating-twitter7

end server that dispatches the query as well as user preference to other servers. Finally, results from distributed servers are merged and ranked by chronological order. This solution may ensure high efficiency but do not support however advanced ranking algorithms that take into account social and temporal evidences.

Chen et al. (2011a) propose to reduce the size of index by including only high priority tweets from actual trends that have a higher chance to be displayed in search results. In addition to tweet text, the index structure, represented in figure 4.7, maintains the user name, the tweet timestamp, the user's *PageRank* score and the encoded reply tree. This data supports a refined ranking function that considers different factors of relevance in respect of microblog specificity.

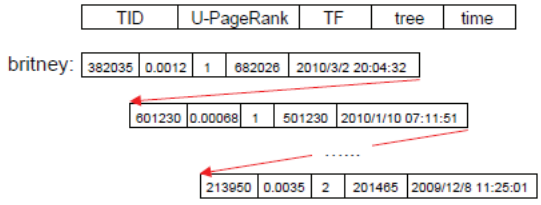


Figure 4.7: Structure of TI Inverted Index (Chen et al., 2011a)

4.3.4 Ranking approaches for microblogs

Despite the problem of microblog indexing which is not widely discussed in the literature, a wide range of works have focused however on microblog search and tweet ranking. This problem presents open challenges due the diversity of search motivations and relevance features.

The two main questions that have been raised in this context are: *"Which features reflect better the relevance of tweets?"* and *"how to combine them?"*. The answer to the first question remains unsolved due the variety of relevance features as presented in section 4.3.2. Some empirical studies such as (Duan et al., 2010; Tao et al., 2012) show critical experiments problems of scalability and biases. Regarding the second question that addresses instead the combination of relevance factors, we identify next four main approaches.

Linear combination approaches. This category of approaches combines using a linear function different relevance indicators. Chen et al. (2011a) propose to combine two factors that model the on the first hand the relevance of the tweet and on the second hand the relevance of respective reply tree. The first part depends in fact on the similarly of the tweet to the query $sim(q, t)$, the authority of the microblogger in the followers' social network *U-PageRank*, and the time

decay between the query and the tweet ($q.timestamp - t.timestamp$). The second part depends on the popularity of the involved microbloggers in the discussion topic $tree.popularity$ and discussion time ($q.timestamp - tree.timestamp$). These features are combined as defined in the next equation:

$$\mathcal{F}(q,t) = \frac{w_1 \times U - PageRank + w_2 \times sim(q,t)}{q.timestamp - t.timestamp} \tag{4.5}$$

$$+ \frac{w_3 \times tree.popularity}{q.timestamp - tree.timestamp} \tag{4.6}$$

where w_1, w_2 and w_3 are tree weighting parameters that enable to emphasize respectively, the microblogger authority, the tweet similarly with regard to the query and the popularity of the discussion thread. .

In the same approach, Nagmoti et al. (2010) propose a linear combination function of three relevance features including the fraction of microbloggers' followers $f_{FR}(t,q)$, the tweet length $f_{LR}(t,q)$ and the presence of URLs $f_{UR}(t,q)$. The first score $f_{FR}(t,q)$ models the social relevance of the tweets. The second and the third scores, $f_{LR}(t,q)$ and $f_{UR}(t,q)$, evaluate the quality of the tweets. These scores as combined as defined next:

$$f_{FLUR}(t,q) = f_{FR}(t,q) + f_{LR}(t,q) + f_{UR}(t,q) \tag{4.7}$$

Machine learning approaches. This category of approaches uses a machine learning algorithm in order to combine the relevance features. First, a set of features scores may compute for each tweet. These scores represent the relevance of the tweet into a multidimensional space. Based on a training dataset, a learning to rank algorithm is after that applied on the result set in order identify top relevant tweets. Despite the representations approach (*e.g.*, pointwise (Nallapati, 2004), pairwise (Herbrich et al., 1999)), *etc*) and the ranking algorithms (*e.g.*, RankBoost, SVM (Herbrich et al., 1999), (Freund et al., 2003), *etc*), the core component of learning to rank models for microblog search consist on the includes features.

Duan et al. (2010) propose a learning to rank approach that uses three types of features. Content relevance features investigate tweet content (*BM25 score*, *Similarity of contents*, *Length*). Twitter specific features evaluate tweet quality (*URL*, *Retweets*, *hashtags*, *replies*). Account authority features evaluate the tweet author. The main score in this category *Popularity Score* is computed by applying PageRank algorithm on the social network of retweets.

Metzler and Cai (2011) propose a learning to rank approach that considers the textual similarity between the query and the tweet (*text score*), the time difference between the query and the tweet (*tdiff*), the hashtag existence (*has hashtag*), the URL presence (*has url*), the percentage of words out of vocabulary (OOV) and the tweet length (*length*).

Unified approaches. This category of approaches integrates several relevance features into a unified framework. For instance, Liang et al. (2012) propose a real-time tweet ranking model based on the language model framework. The probability of generating the query Q given a tweet D and a timestamp t is de defined by:

$$P(Q|D,t) = P(t|Q,D) \sum_{w \in v} P(w|\hat{\theta}_Q) log P(w|\hat{\theta}_D) \tag{4.8}$$

where $\hat{\theta}_Q$ denotes the query model, $\hat{\theta}_D$ is the document model and $P(t|Q,D)$ defines the temporal re-ranking component.

Probability $P(t|Q,D)$ is computed based the document's temporal profile $\mathcal{N}$ as follows:

$$P(t|Q,D) = e^{-\frac{\mathcal{N}}{k}} \tag{4.9}$$

where k is an exponential rate parameter, $\mathcal{N} = (t^* - t_D)/H$, t_D is the tweet timestamp, t^* is the query timestamp and H is a normalizing interval factor.

Relevance feedback approaches. This category of approaches proposes to expand the query with additional terms. In fact, both query and tweets are short. Adding more representative terms to the query helps to gather more tweets in the topic. In particular Liang et al. (2011) propose to expand the query using The New York Times news headlines, WorNet and the subset of top 100 tweets relevant of tweets, presumed in this context relevant. The relevance score of the tweet is defined by the following equation:

$$P(Q|T) = \prod_{i=1}^{n} P(q_i|T) \tag{4.10}$$

$$P(q_i|T) = (1-\lambda)\frac{f_{q_i,T}}{|T|} + \lambda \frac{C_{q_i}}{|C|} \tag{4.11}$$

where $f_{q_i,T}$ is the frequency of term q_i in the tweet T, f_{C_i} is the frequency of term q_i in the subset of 100 top tweets.

Bandyopadhyay et al. (2011) propose to expand the query based on the title of Web search results. Most frequent words or n-grames over the top search results of Google Search API are added to the query.

In the same approach, Li et al. (2011b) propose to extract the words with a strong connection to the topic in order to expand the query. Term similarity is estimated in this case based on the term association network and the term resistance network

4.4 Identifying influencers in microblogs

The intensive microblogging activities have bring forward a group of microbloggers with a prominent position in the social network. They play a key role in their social neighborhood as well as at the entire social network level. In microblogging context, these microbloggers are known as network *"influencers"*. The term *"influencer"* covers indeed several properties that assert the social relevance of a person. Bakshy et al. (2011) defined influencers as users who exhibit a combination of personal and social network desirable attributes.

> *influencers [...] exhibit some combination of desirable attributes — whether personal attributes like credibility, expertise, or enthusiasm, or network attributes such as connectivity or centrality —that allows them to influence a disproportionately large number of others.*

This section focuses on the problem of indentifying microblog influencers. We present in what follows an overview of main influencer attributes and ranking algorithms proposed in the literature. Before discussing this, we will present the main social network models for microblogging networks.

4.4.1 Microblog social network

In order to represent the social network of microblogs, several work (Kwak et al., 2010; Weng et al., 2010; Lim and Datta, 2012) propose to use native social relationships explicitly defined by microbloggers, *i.e.*, followership. The followers social network connects microbloggers with respective followings in the network, and conversely, to their followers. Using graph formalism, followers social network is represented by a directed graph $G(V, E)$.

Definition 6 (Follower network) *A directed graph $G(V, E)$ is formed with the twitterers and the "following" relationships among therm. V is the vertex set, which contains all the twitterers. E is the edge set. There is an edge between two twitterers if there is "following" relationship between them, and the edge is directed from follwer to friend (Weng et al., 2010).*

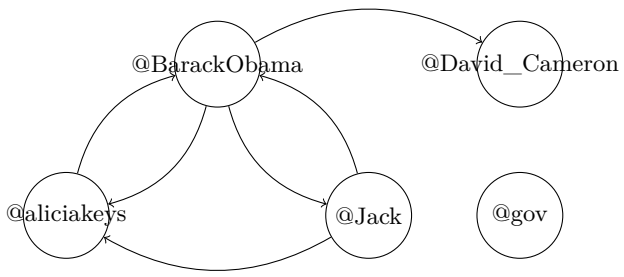


Figure 4.8: Followers social network of popular of Twitter users

Figure 4.8 illustrates an example of followers social network from the list of popular tweets of United States presidential election previously introduced in figure 4.1. This social network is extracted using Twitter social graph.

In addition to followers social network, microbloggers are represented using retweet relationships (Duan et al., 2010; Conover et al., 2011; Greene et al., 2011; Ota et al., 2012). Retweets that connect in particular tweets are transferred to related microbloggers. Retweet relationships reflects in this context endorsement between users (Boyd et al., 2010). Similarly to the followers network, retweet social network is modeled by a directed graph.

Definition 7 (Retweet network) *In the retweet network an edge runs from a node representing user A to a node representing user B if B retweets content originally broadcast by A, indicating that information has propagated from A to B (Conover et al., 2011).*

Reply feature enables users to post comments on microblogs. A discussion tree is built by joining tweets and replies. Several works in literature (Sousa et al., 2010; Khrabrov and Cybenko, 2010; **?**; Rossi and Magnani, 2012) propose to use discussion tree in order to represent the social network of users. Accordingly, a reply relationship is defined between users involved in the discussion. Such social network representation models communication aspects among microblogs. Reply network, known also as discussion network and conversation network, is represented with a graph $G(V, E)$.

Definition 8 (Reply network) *Such implicit network derived from tweet replies can be represented as a directed graph $G = (V, E)$ where each node $u \in V$ represents a user and each edge $(u_i, u_j) \in E$ represents an @reply message sent from user u_i to user u_j (Sousa et al., 2010).*

One alternative to model discussion within microblogs is to use mention network (Yang and Counts, 2010; Conover et al., 2011). Mentions are indeed a generalized form of discussion where the tweet is addressed to one or more microbloggers. Replies are specific tweets that mention the microblogger of answer post. Similarly to previous social network models, mention networks is represented with a directed graph.

Definition 9 (Mention network) *In the mention network, an edge runs from A to B if A mentions B in a tweet, indicating that information may have propagated from A to B (a tweet mentioning B is visible in B's timeline) (Conover et al., 2011).*

We notice that native retweets are transformed in text format as *"RT @username"*. This implies a new mention relationship defined between users even though the original purpose was a retweet. Nevertheless, unofficial retweets starting also with *"RT @username"* are considered as simple mentions.

.

4.4.2 Measuring influence on microblogs

Although microblog influence has been widely discussed in literature, there still no comprehensive definition that characterize this property. Several assumptions are introduced in this purpose in order to understand influence and to provide objective measures that evaluate it. We discuss in what follows the main microblog influence assumptions that have been addressed in the literature.

Popularity. Influence is interpreted as the popularity of the microblogger in the social network. The more popular a person is, the largest visibility respective tweets will gain in the social network. Different metrics are proposed in this context to quantify the popularity of a microblogger. Several work propose to use followership count as an indicator of popularity. For instance, (Kwak et al., 2010) propose to measure the popularity based on the number of followers. However, this measure is exposed to spam problem as microbloggers have the ability to increase the number of their followers thanks to some promotional methods such as account advertising, followers recruiting and the *"follow me, I follow you"* practice. To resolve this problem, Nagmoti et al. (2010) propose to consider the number of followings which reflect the social activity of the microblogger. Accordingly, popular microblogger must, simultaneously, follow and be followed by many users. In particular, *FollowRank* measures the popularity of microbloggers as the proportion of followers over the total number of followers and followings as defined in the next equation.

$$FR(a) = \frac{i(a)}{i(a) + o(a)} \tag{4.12}$$

with $i(a)$ is the number of followers of microblogger a and $o(a)$ is the number of followings of microblogger a.

Authority. In addition to popularity, some works have addressed influence within microblog as a matter of authority in the social network. Link analysis algorithms such as *PageRank* and *HITS* are used in this context to indentify authoritative people in the follower network (Kwak et al., 2010; Gayo-Avello and Brenes, 2010). Contrary to popularity based influence, this methodology is less sensitive to spam problems. In fact, such approaches overcome the problem of excessive microblogging activity that aims to generate more followers namely by publishing too many tweet, random retweets and pointless replies. Inspired by *PageRank*, works in literature propose to investigate mutual authorities in the followers social network. In particular, *TunkRank*[8], one of the first measures proposed in this context (2009), apply a slight modification on original *PageRank* algorithm in order to boost microbloggers with a low number of followers.

[8] http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/

TunkRank is computed using next formula.

$$Influence(X) = \sum_{Y \in Followers(X)} \frac{1 + p \times Followers(Y)}{\|Following(Y)\|} \tag{4.13}$$

with $Followers(X)$ corresponds to the set of followers of microblogger X and, conversely, $Followings(Y)$ corresponds to the set of followings of microblogger Y.

We notice that original *PageRank* algorithm is also applicable in this context in order to identify authorities in followers social network, retweet social network and mention social network (Kwak et al., 2010; Duan et al., 2010). For this aim, microblogger are treated as similar to Web pages. Social relationships including followerships, retweets and mentions are considered here as hyperlinks.

In order to identify topical authorities in the social network, Weng et al. (2010) propose to apply topic-sensitive PageRank on followers social network. In particular, their proposed *TwitterRank* algorithm is computed iteratively using the next equation:

$$\vec{TR}_t = \gamma P_t \times \vec{TR}_t + (1 - \gamma) E_t \tag{4.14}$$

where P_t is the transition probability matrix that models the topical similarly between two microbloggers in topic t, E_t is the teleportation vector of the random surfer in topic t and $\gamma \in [0, 1]$ is the random surfer probability as defined in original *PageRank* algorithm.

Information diffusion. Beyond these previous assumptions, some works define influence as the ability to spread information through the social network. Retweet count is proposed in this context as a basic metric for information diffusion (Kwak et al., 2010). Obviously, *PageRank* score computed on the retweet social network Duan et al. (2010) may be used instead to identify mutually retweeted microbloggers.

From another perspective of view, Bakshy et al. (2011) propose to investigate the propagation of *"seed"* URLs in the microblogging network. An influencer is thus highlighted if he introduces a new URL that has been subsequently reposted by other users either by means of retweets or through regular tweets. To illustrate that, let A and B be two microbloggers with microblogger B is following microblogger A. We assume that microblogger A has influenced microblogger B if A has published, at time t, a URL that was reposted later, at time $t+\epsilon$, by microblogger B. In spite of retweet count that focuses on retweets, this approach investigates also information diffusion within regular tweets. We notice that access to followership data is compulsory to perform this approach. Nevertheless, this is not always available due to privacy and scalability issues.

Conversational. Influence microblogging network is interpreted as the capability to initiate conversations and engage audience (Cha et al., 2010). Accordingly, the influence of user is measured by the number of received mentions. Experimental results shows that most mentioned users correspond to celebrities (Cha et al., 2010). These users receive much attention in the social network. In order to measure the mention impact, Pal and Counts (2011) propose to consider the number of outgoing mentions ($M1$), the number of mentioned users ($M2$), the number of received mentions ($M1$) and the number of mentioning users ($M4$). The mention impact MI is defined by the next equation:

$$MI = M3 \cdot \log(M4) - M1 \cdot \log(M2) \tag{4.15}$$

To evaluate the conversional activities of a microblogger typically the ability to receive comments on his tweets, Pal and Counts (2011) propose to measure the non-chat signal CS of the microblogger as defined next:

$$CS = \frac{OT1}{OT1 + CT1} + \lambda \frac{CT1 - CT2}{CT1 + 1} \tag{4.16}$$

With $OT1$ is the number of original tweets, $CT1$ is the number of conversational tweets, $CT2$ is the number of conversational tweets initiated by the microblogger and λ is tuning parameter. Setting λ to large values enables to highlight microbloggers with intensive social interactions.

Some work address microblog influence as a composite property of microblog actors that combines several features. Cha et al. (2010) define influence as the combination of microblogger popularity (Indegree influence), information diffusion (Retweet influence) and the ability to engage other people in conversation (Mention influence). Pal and Counts (2011) define influence as the combination of topical signal (TS), signal strength (SS), non-chat signal (CS), retweet impact (RI), mention impact (MI), information diffusion (ID) and Network score (NS). These features are combined using probabilistic clustering model.

4.5 TREC Microblog track

TREC Microblog is an evaluation campaign for microblog retrieval organized annually since 2011 in conjunction with TREC workshop. The goal of this track is the join research community that interest in microblogs and design an evaluation protocol to microblog retrieval systems. TREC Microblog includes a main adhoc task, known as *real-time adhoc search*, and a second filtering track introduced in 2012. Both tasks are based on *tweet2011* corpus. This dataset includes about 16 million tweets published over 16 days (January, 17_{th} - February, 2_{nd} 2011). The dataset is built based on public Twitter Stream API which provides a representative sample of 1% of the tweet stream.

The goal of real-time ad-hoc search task is to find most relevant tweets and also the most recent tweets for a query q given a date. For this purpose, a set of time-stamped topics are determined by track organizers. 2011 queries dataset include 49 topics while 2011 topic dataset include 60 queries. A sample topic is represented in figure 4.9. Query time represents the time where the query is submitted and querytweettime corresponds to the identifier of the last tweet submitted before the query which is helpful to filter tweets instead of comparing date.

```
<top>
  <num>MB01</num>
  <title>Wael Ghonim</title>
  <querytime>25th February 2011 04:00:00 +0000</querytime>
  <querytweettime>3857291841983981</querytweettime>
</top>
```

Figure 4.9: TREC Microblog topic sample

The relevance judgments are constructed from top 30 results of submitted systems. In Microblog 2012 track, the pool depth was extended to top 100 results of each system. A group of NIST assessors has assed tweets and assigned a gradual relevance between -2 and 2 (-2: Spam; 0: Not Relevant, 1: Minimally Relevant, 2: Highly Relevant). During relevance judgment, only tweets in English are analyzed while retweets are automatically considered irrelevant. To respect the time constraint, no source of information posterior to the query may be used including tweets in the collection published after the date of the query. The final classification results are established in the reverse chronological order unlike other TREC tasks that classify documents according to their score.

In 2011, results are evaluated based on the precision at rank 30 (P@30). Mean average precision MAP is used as non official measure for a deep analysis of submitted run. These measures were replaced in 201 with P@30 ranked by score and the ROC curve (Fawcett, 2006).

Even though this task has received a lot of success in terms of number of participants (largest TREC task), some revision need to be applied on the corpus and evaluations measures. For 2013 edition, the organizers plan to release a new corpus which covers two months of Twitter stream (about 260 million tweets) and also to transform this track into a service oriented task where the dataset is accessible via a search API and retrieval processes are conducted online.

Conclusion

We presented in this chapter main social approaches for microblog retrieval. In particular, we discussed the search motivations and the relevance factors of microblog search task and we gave an overview of proposed models in this context. Moreover, we focused on the problem of microblogger ranking and we presented the main approaches for identifying key microblogs in the social network, namely network influencers. Finally, we presented in this chapter the TREC microblog Track.

In accordance to the main subjects discussed in this chapter, namely microbloggers ad microblog ranking, we will propose in chapter 6 et 7 two social models for ranking tweets and microbloggers in microblogging social network.

Part II

Ranking social relevance in information networks

Chapter 5

A social information model for flexible literature access

5.1 Introduction

Information retrieval within bibliographic resources differs from other application domains by a specific information need of users who require a high scientific quality of retrieved documents. As a consequence, scientific indexers and academic digital libraries have addressed one common issue: evaluating the scientific quality of bibliographic resources. To tackle this problem, literature retrieval approaches have integrated scientific impact metrics as a key feature for ranking retrieval results.

In a boarder context, documents and authors are inseparable entities and may represent each other. The quality of a document indicates thus the quality of related authors and vice versa. Based on this idea, recent approaches in literature retrieval addressed the social network of bibliographic resources and evaluated their quality by exploiting the social importance of the corresponding authors. Co-authorship networks are commonly used in this area of research to represent the social context of bibliographic resources (Yan and Ding, 2009; Mutschke, 2003).

With the introduction of scholarly social bookmarking services, the importance of scientific documents is not only inferred from their production context but also using the social consuming context. As illustrated in figure 5.1, the social network of bibliographic resources involves several entities that interact in the social producing and consuming contexts. In addition to documents, these entities include information producers (e.g. authors), information consumers (e.g. users, annotators) and social annotations (e.g., tags, rating, reviews). Accordingly, the importance of the scientific publications is estimated based on related

entities particularly the social importance of related actors (Amer-Yahia et al., 2007).

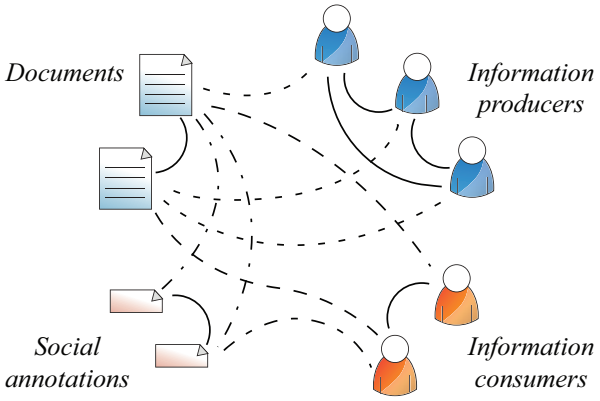


Figure 5.1: Social producing and consuming context

With this in mind and inspired by the work of Kirsch et al. (2006) and Mutschke (2003) representing bibliographic resources with social networks, we introduce a social information retrieval model for literature access that evaluates the social relevance of scientific publications. For this aim, we propose to combine the topical relevance of documents and the social importance of respective authors and annotators in the social network. The two main issues addressed here are the modeling of bibliographic social networks and to the evaluation of document relevance based on the position of respective actors in the social network. In the contrast of closely related works (Kirsch et al., 2006; Mutschke, 2003; Kirchhoff et al., 2008), our model presents the following features:

- Estimating document relevance by the combination of the topical relevance and the social importance of related actors in the social network. In contrast of (Kirsch et al., 2006) using a product function for combing topical and social scores, we propose to combine the topical relevance and the social importance of authors using a linear combination;
- Including citations links as social interactions between authors of scientific papers unlike work of Kirsch et al. (2006) and Mutschke (2003) which only use co-author relationships for modeling the social network of authors.
- Defining a weighting model for edges connecting social entities to evaluate influence, knowledge transfer and shared interest between authors in the contrast of approaches in (Kirchhoff et al., 2008) and (Mutschke, 2003) using basically a binary network model.

Regarding our previous work (Ben Jabeur et al., 2010; Ben Jabeur and Tamine, 2010), new contributions presented here are:

- Taking into account the joint authors network and the network of annotators in the calculation of the overall relevance of the documents;

– A new measure of the social importance of the authors and annotators of bibliographic resources, based on the quantification of their expertise;
– An experimental evaluation based on a new standard corpus, namely CiteData (Harpale et al., 2010) which available to the scientific community since 2012.

The rest of this chapter is organized as follows. First, we introduce the literature information network then we define the qualitative and the quantitative models of authors' social network as well as users' social network. Then, we present the global approach for combining topical and social relevance. Afterward, we will focus on the evolution of the social importance of scientific authors. Finally, we conduct a series of experiments based on CiteData corpus in order to evaluate the effectiveness of our model.

5.2 Literature information network

An information network is a graph-based representation of information units as well as dependency, structural and semantic relationships between them. Likewise, a social network is a graph-based representation of individuals and possible social relationships between them. By integrating both representations, the social information network provides a unified network model that represents data (information units), actors (individuals) and their mutual interactions

Citation network is a common representation of literature's information network. It represents documents with reference relationships. The bipartite graph of tags and articles is another example of information network in a scholarly social bookmarking environment. As discussed in section 3.1, authors of bibliographic resources are represented using several social network models typically co-authorship and author citation network. In this work, we propose to combine all these network models in order to represent actors and data in interaction with bibliographic resources during the producing and consuming processes. In particular, we model bibliographic resources based on authorship, citation and social bookmarking. The proposed social information network includes two types of data entities namely documents and tags and two types of actors including authors and social bookmarking users.

Based on graph notation, the social information of bibliographic resources is represented by a graph $G = (V, E)$. The set of nodes $V = A \cup U \cup D \cup T$ denotes actors and information entities including authors A, social bookmarking users U, documents D and tags T. The set of edges $E \subseteq V \times V$ represents the different relationships between entities. The main relationships of the social information network of bibliographic resources are identified next:

– *Authorship:* connects an author $a_i \in A$ with his authored document $d_j \in D$.
– *Reference:* connects a document $d_i \in D$ with its referenced documents.

– *Co-authorship:* connects two authors $a_i, a_j \in A$ having produced one common document at least.
– *Citation:* connects two authors $a_i, a_j \in A$ if author a_j cites a_i at least once through his documents.
– *Bookmarking:* connects user $u_i \in U$ and his bookmarked document $d_j \in D$.
– *Annotation:* connects document $d_i \in D$ with tag $t_j \in T$ assigned at least once to describe its content.
– *Tagging:* connects user $u_i \in U$ and tag $t_j \in T$ since he uses it at least once to bookmark a document.
– *Friendship:* connects users $u_i, u_j \in U$ if either they have a direct personal relationship or they join the same group.

Figure 5.2 illustrates an example of social information network. We note that network nodes and relationships are explicit and immediately available through literature and social bookmarking databases without further extraction process.

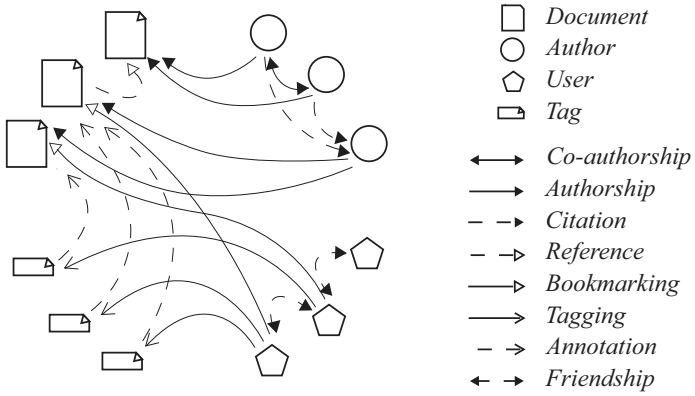


Figure 5.2: The social information network of bibliographic ressources

5.3 The Social network model

In order to evaluate the social relevance of a document we define two sub-social networks that represent authors and users. The two social networks models describe the social producing context and the social consuming context of bibliographic resources.

5.3.1 The social network of authors

The scientific publication process is characterized with two main aspects: collaboration and knowledge transfer. In particular, collaboration is realized through

the co-publication of scientific research. Co-authorship ties reflect in this context the collaboration structure within the scientific community. Knowledge transfer is realized through citation practices where referenced articles have introduced valuable findings that inspire authors in their work. In order to represent collaboration and knowledge transfer, we propose to model the social network with co-authorships and citation relationships. We present in what follows the qualitative and quantitative components of the social retrieval model.

5.3.1.1 The qualitative model component

The social network of authors is represented by a directed and weighted multigraph $G_A = (A, E)$ where the set of nodes A represent authors and the set of edges $E \subseteq A \times A$ represents the social relationships between them, namely co-authorship and citation.

Co-authorship is basically a collegial relationship that equally connects each couple of co-authors a_i and a_j having co-published at least one common article. In order to transform this undirected association into a directed relationship, two reciprocal edges $(a_i, a_j) \in E$ and $(a_j, a_i) \in E$ are created between each couple of co-authors a_i and a_j in the social network.

Citation is an implicate relationship primarily defined on articles and transmitted to authors by means of authorship associations. Accordingly, a citation relationship is defined from author a_i to a_j if a_i cites in his publications an article authored by a_j .

We notice that a couple of authors may be connected with different types of edges at the same time particularly if an author cites one of his co-authors. Unlike simple graph, this property is supported by multi-graphs model of the social network where parallel edges are allowed between the same pair of source and target.

5.3.1.2 The quantitative model component

As stated above, the social network of authors includes different types of relationships. Although these social relations are represented alike in a social network, they express different semantic. We propose to assign representative weights to each type of relationship with regard to their importance.

Co-authorship expresses similarly and a shared interest between authors. These properties are reinforced as authors collaborate with several co-authors in common. Accordingly, we propose to weight the network edges according to the

social relationship type.

$$w(i,j) = \frac{Co(i,j)}{Co(i)} \tag{5.1}$$

Where $Co(i,j)$ is the number of common co-authors of a_i and a_j. $Co(i)$ is the number of co-authors of a_i. We notice that edges (a_i, a_j) and (a_j, a_i) have different weights in respect of co-author number.

Citation relationship expresses knowledge transfer between authors. As an author often cites articles of a particular one, he would be influenced by his work and discuss related topics. Accordingly, we propose to weight citation edges based in respect of citation number as defined next:

$$w(i,j) = \frac{Ci(i,j)}{Ci(i)} \tag{5.2}$$

where $Ci(i,j)$ is the number of times a_i cites a_j and Ci_i is the number of citations expressed by a_i.

5.3.2 The social network of users

The social network of users is represented by a directed and weighted graph $G_U = (A, E)$ where the set of nodes U represents social bookmarking users and the set of edges $E \subseteq U \times U$ represents friendship between them.

Once friendship feature are not often available in scholarly bookmarking service, we propose to infer these relationships from group membership and co-tagging activities.

The interest of been subscribed to a bookmarking group is to be notified about the recent tagging activities in some predefined topic. Hence, users of the same group implicitly claim shared interest with each other. Let G be the set of a social bookmarking groups. Two users u_i and u_j are called friends if they join at least one common group $g_i \in G$, in another word, $u_i, u_j \in g_i$. In this case, two symmetric edges with equal weights are defined between the two microbloggers $(u_i, u_j) \in E$ and $(u_j, u_i) \in E$. Friendships weights $w(i,j)$ and $w(j,i)$ are assigned respectively to (u_i, u_j) and (u_j, u_i) as defined next:

$$w(i,j) = w(j,i) = 1 \tag{5.3}$$

Besides group membership, user friendship could be inferred from co-tagging activities. Let T_{u_i} and T_{u_j} respectively be the set of papers tagged by user u_i and the set of papers tagged by user u_j. Users u_i and u_j are called friends if both of them have tagged at least one common paper, in another word, $T_{u_i} \cap T_{u_j} \neq \varnothing$.

Friendship weights are defined in this case with respect to edge direction as defined next:

$$w(i,j) = \frac{|T_{u_i} \cap T_{u_j}|}{|T_{u_i}|} \tag{5.4}$$

We notice that friendship weight $w(i,j)$ is set to 1 if u_i and u_j are subscribed to the same group regardless of their co-tagging activities.

5.4 Social importance of authors and users

In order to evaluate the social importance of scientific authors and users of scholarly bookmarking networks, we propose a new algorithm called *SoRank* that identifies important actors in bibliographic social networks. In particular, *SoRank* is a *PageRank* like algorithm that considers, in addition to the social network structure, the expertise of actors on a particular topic. In order to estimate the expertise of an actor on a particular topic, represented in our case by a query q, we first compute the relevance score of each document using the language model (Hiemstra, 2001).

$$P(d_j|q) = \sum_{t_i \in q} \log \left(1 + \frac{\lambda t f_{t_i,d_j} \sum_t c f_t}{(1-\lambda) c f_{t_i} \sum_t t f_{t,d_j}} \right) \tag{5.5}$$

An expertise score is then assigned to each actor as the sum of relevance scores $P(d_j, q)$ of related documents $\mathscr{D}(r_i)$. To compute an expertise score in the range of $[0, 1]$, actor's score is normalized by the division by the sum of relevance score of all documents as follows.

$$w(r_i) = \frac{\sum_{d_j \in \mathscr{D}(r_i)} P(d_j|q)}{\sum_{d_k \in C} P(d_k|q)} \tag{5.6}$$

where $\mathcal{D}(r_i)$ corresponds to the set of published documents in the case of author social network and to the set of the tagged documents in the case of user social network structure. C represents the document collection.

The expertise of actors represents the probability of selecting a random actor having a query q. On the other hand, the probability to move from an actor to another one depends on the structure of the social network. Accordingly, a social importance score is assigned to each actor in the social network as defined next.

$$SoRank^t(r_i, q) = \frac{(1-d)w(r_i)}{\sum_r P(r|q)} + d \sum_{r_k \in \mathscr{P}(r_k)} w(r_k, r_s) \frac{SoRank^{t-1}(r_k)}{|\mathscr{S}(r_k)|} \tag{5.7}$$

where $d \in [0, 1]$ is a damping factor as similar to PageRank algorithm, $\mathscr{P}(r_k)$ the set predecessors of node r_k and $\mathscr{S}(r_k)$ the set successors of node r_k. Relationship weight $w(r_k, r_s)$ is computed according to the type of the social relationship as defined previously in section 5.3.1 and 5.3.2.

SoRank is applicable on the social network of authors and the social network of users. It helps to identify authoritative experts in the social network. These actors correspond to authoritative authors in the social network collaborating and cited by experts authors too. Likewise, important actors in the user social network correspond to central users in the social network endorsed by similar actors.

SoRank propagates expertise scores of the network actors through incoming edges with respect of the relationship weights. This process is repeated iteratively until ranking convergence. For each iteration, scores are normalized by division by the sum of all the nodes. The detailed descriptions of SoRank algorithm is presented next.

Algorithm 1: SoRank

$t \longleftarrow 0$
foreach $u_i \in U$ **do** $SoRank^t(r_i, q) = 0$; `// initialization`
repeat
  foreach $r_i \in R$ **do**
    $SoRank^t(r_i, q) = \frac{(1-d)w(r_i)}{\sum_r w(r)} + d \sum_{r_k \in \mathscr{P}(r_s)} w(r_k, r_s) \frac{SoRank^{t-1}(r_k)}{|\mathscr{S}(r_k)|}$
  end
  foreach $r_i \in R$ **do** $SoRank^t(u_i) = \frac{SoRank^t(u_i)}{\sum_r SoRank^t(r)}$
  $t \longleftarrow t + 1$
until *convergence*

Convergence is assumed whenever node ranking remains the same for two consecutive iterations.

$$\forall r \in R, \quad Rank^t(r) = Rank^{t-1}(r) \tag{5.8}$$

where $Rank^t(r)$ and $Rank^{t-1}(r)$ correspond to the rank of node r at iterations t and $t-1$, respectively.

By setting expertise score $w_r = 1$ for all the actors in the network, *SoRank* produces a topical-independent rankings of actors. In this case, *SoRank* is assimilated to a weighted version of *PageRank*.

5.5 Combining topical and social relevance

In order to enhance the retrieval process within social bibliographic resources, we propose to combine the social network structure into the ranking process involved during document retrieval process. For this aim, we propose a modular approach for social information retrieval that combines (a) the topical relevance and (b) the social relevance of related actors as illustrated in figure 5.3. The topical relevance is based on the similarity between query and document. The social relevance depends on the position of related actors in the social network, typically on the expertise and the authority of actors and users as defined previously by *SoRank* algorithm. By combining topical and social relevance factors, retrieval process identify high quality of documents discussing the query topic and also in relation with important actors in the social network.

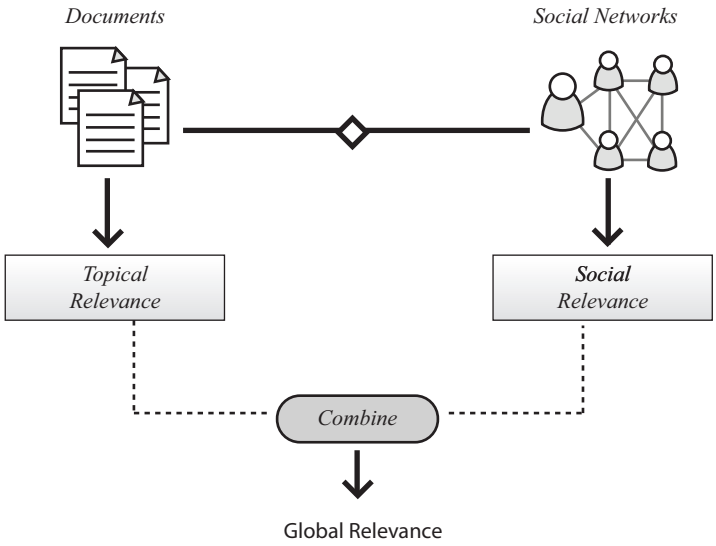


Figure 5.3: A modular approach for Social Information Retrieval

The social information retrieval model for literature access estimates for each document d a global relevance score $Rel(d, q, G)$ that considers the query q, the document q as well as the social network of related actors G. In fact, the $Rel(d, q, G)$ combines two relevance features namely the topical relevance, represented by $RSV(q, d)$ and the social relevance $S(d, qG)$. A relevance score is computed for each feature then computed as follows:

$$Rel(d, Q, G) = \alpha RSV(d, Q) + (1 - \alpha)S_d(d, Q, G) \tag{5.9}$$

where $\alpha \in [0, 1]$ is a tuning parameter that balances between the topical relevance and the social relevance of documents. By setting α to 1, only the topical

relevance score is considered. In this case, $Rel(d, q, G)$ produces similar rankings as $RSV(d, q)$. With $\alpha = 0$, only the social relevance $S = (d, q, G)$ is taken into account. In this case, documents are ranked based on the social relevance of related actors. The α parameter depends on the user needs, search task and query topic sensibility. We conduct in section 5.6 empirical experiments in order to select appropriate values of α for ad-hoc search task within bibliographic resources. We notice that topical and social scores are normalized as defined next.

$$f(s_i) = \frac{s_i - min(s)}{max(s) - min(s)} \tag{5.10}$$

where $min(s)$ and $max(s)$ are minimum and maximum values for each relevance feature score.

We detail in what follows topical relevance and social relevance involved in our model.

5.5.1 Topical relevance

The topical relevance determines the topical overlap between the query and the document. This is estimated through term occurrence in query and documents as defined by traditional information retrieval models. In this context, probabilistic model Okapi BM25 (Robertson et al., 1995) model is used to compute a topical relevance score for each document.

$$RSV(d, q) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i, d)(k_1 + 1)}{f(q_1, d) + k_1.(1 - b + b\frac{|d|}{avgdl})} \tag{5.11}$$

where $f(q_i, d)$ denotes the frequency of term q_i in document d, $|d|$ represents the document length and $avgdl$ the average document length in the collection. Inverse document frequency of query term q_i is computed as follows:

$$IDF(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{5.12}$$

with N is the collection size and $n(q_i)$ represents the number of documents containing query term q_i.

Other traditional retrieval models can be used instead of the Okapie BM25 to estimate the topical relevance of documents. Meanwhile, the used model may impact the result effectiveness as results set are built from retrieval results then ranked by combining the initial topical relevance with the social relevance score.

5.5.2 Social relevance

The social relevance score determines the social importance of related actors in the social network. Although a social relevance score $S(d, q, G)$ is assigned

to each document, this score is computed for each related actor. However, a document may be related to many actors from the same type typically several co-authors or different users that have tagged the document. To handle this issue, an aggregated social score of related actors is computed by combining authors and users social score that better express the social relevance of documents in the social producing and consuming context.

The social relevance score is computed by aggregating expert authority scores as expressed by *SoRank* algorithm introduced in section 5.4. In particular, a document social score is computed as the sum of SoRank scores of related actors $\mathscr{A}(r_i)$ as defined by the next equation.

$$S'(d, q, G) = \sum_{r_i \in \mathscr{A}(r_i)} SoRank(r_i, q) \tag{5.13}$$

where $\mathscr{A}(r_i)$ correspond to the set of document co-authors in the case of author social network and to the set of the social bookmarking users that have tagged the document in the case of user social network.

In practice, two social relevance scores $S(d, q, G_A)$ and $S(d, q, G_U)$ are computed to each document considering the social network of authors and the network of users. Both score reflect the social relevance of the document but in different social contexts. A final score is then concluded by selecting the maximum score using *CombMax* operator. Accordingly, documents receive a high relevance score if either corresponding authors or users are important in the social network. The social relevance score of a document is defined by:

$$S_d(d, q, G) = CombMax\left(S'(d, q, G_A), S'(d, q, G_U)\right) \tag{5.14}$$

Obviously, other combination operators can be used to selected target social network typically the *sum* of both social importance scores and a weighted *sum*. *CombMax* operator leverages however the social importance score for recently paper that have not yet enough time to be cited but tagged by different users.

5.6 Experimental results

In order to validate our social information model for literature access, we conduct a series of experiments on a dataset of scientific publications including about 78000 articles. In particular, we evaluate the impact of social network features for ranking literature in an ad-hoc retrieval task. As users of this information retrieval task are interested in topically relevant articles but also related to important actors in the social network, these experiments study both relevance factors of literature ranking and retrieval namely the topical relevance and the social relevance. The main goals of these experiments are:

– Study the impact of the social network structure in particular the social network of authors and scholarly bookmarking users.

– Compare the effectiveness of scientific impact and social network measures for ranking literature.
– Evaluate the retrieval effectiveness of our social retrieval model.

5.6.1 Experimental setup

In order to evaluate an ad-hoc retrieval task over literature data, we used in our experiments the multifaceted dataset of scientific articles CiteData (Harpale et al., 2010). The available distribution of the dataset includes titles, abstracts and citation network of about 81000 articles but no authorship and social bookmarking data. We collected missed information from recent CiteSeerX and CiteULike dumps. We present in what follows a detailed description of the collected dataset, evaluation measures and the compared literature retrieval models.

5.6.1.1 Article and query dataset

CiteData is a test collection for personalized information retrieval including 81432 academic articles extracted from CiteSeerX and CiteULike repositories. Articles in the dataset cover mainly 11 topics in computer science research field. Table 5.1 and figure 5.4 present the list of topics and respective distributions in the dataset. In addition to articles, CiteData includes 9 search tasks with an average of 5 queries per task defined by domain experts. An example of CiteData search task is presented in table 5.2. About 1936 documents on average were annotated by the same experts in order to identify relevant documents to their need. As we are interested in ad-hoc retrieval task, queries are treated independently. Only relevance annotations for the considered queries are used to evaluate ad-hoc results regardless of search preferences and expert annotations within the same search task.

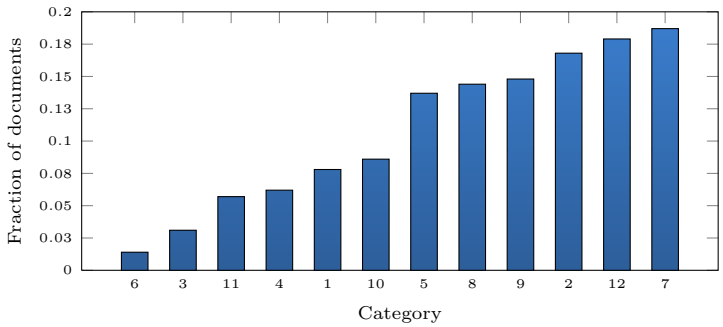


Figure 5.4: Topic distribution of the CiteData dataset (Harpale et al., 2010)

ID	Topic
1	Computer Programming
2	Machine Learning and AI
3	Networking and Security
4	Computer Architecture
5	Agents and Applications
6	Computer Theory
7	Databases
8	Human Computer Interaction
9	Digital Libraries
10	Web and Information
11	Natural Language Processing
12	Other research areas in Computer Science

Table 5.1: List of available categories in the CiteULike dataset (Harpale et al., 2010)

UserID	network03
Task	Information Network Security
Task	Statement Access control is the process in which a request to a data resource or service is mediated to determine whether the access should be granted or denied...
Query1	role based access control
Query2	work flow access control
Query3	authorization delegation
Query4	distributed access control
Query5	XML access control

Table 5.2: Search Task "Information Network Security" (Harpale et al., 2010)

The public available distribution of CiteData includes only titles, abstracts and citation network of articles. In order to perform our social information retrieval model over this dataset, we collected missed data about authors and bookmarking users from CiteSeerX and CiteUlike datasets. We notice that corresponding CiteSeerX records are extracted by exact title matching. CiteUlike dataset provides however article DOI for efficient matching to other dataset namely CiteSeerX. Table 5.3 presents the size of each source dataset and overlap coefficient between them.

	Citedata		CiteSeerX		CiteUlike	
Citedata	81432	(1.00)	78805	(0.97)	17558	(0.22)
CiteSeerX	78805	(0.97)	1471578	(1.00)	35710	(0.02)
CiteUlike	17558	(0.22)	35710	(0.02)	54230805	(1.00)

Table 5.3: Citedata, CiteSeerX and CiteUlike overlap

The extracted datasets contain 78805 articles with 17558 tagged articles. The number of identified entities and relationships of the social information network is presented in table 5.4. As source collection does not provide unique identifiers for authors, we apply an exact matching on authors' names in order to extract the publication of each author.

Entities		Social relationships			
Articles	130354	Authorship	377444	Bookmarking	33305
Authors	107745	Co-authorship	593195	Tagging	45020
Tags	13821	Citation	11379829	Annotation	58877
Users	1394	Reference	1401503	Friendship	7808

Table 5.4: Social information networks statistics

5.6.1.2 Evaluation measures

Users of ad-hoc retrieval task are mainly interested in top results. Close of half of them examine only the top 20 documents before making a decision (Spink et al., 2001). Thus, $P@20$ is used in these experiments as the primary measure for evaluating the retrieval effectiveness of the proposed model. $P@20$ evaluates the ability of a retrieval model to return relevant documents on the top of 20 results. Moreover, Mean Average Precision (MAP) is used as the second measure to compare the overall precision of the retrieval models.

5.6.1.3 Compared models

We compare in these experiments different models that belong to different information retrieval approaches. Table 5.5 describes the compared models and presents corresponding notations.

BM25	Probabilistic model Okapi BM25 (Robertson et al., 1995)
HiemLM	Language model for information retrieval (Hiemstra, 2001)
Cit	Citation index
Exp-Cit	Expected citation count (equation 3.15)
h-Index	H-index of related author (Hirsch, 2005)
PR	PageRank score of article in citation network
PR-CO	Author PageRank in co-authorship network
PR-Cit	Author PageRank in citation network
Kirsh	Product of topical and author PageRank scores (Kirsch et al., 2006)
SoRank	Our social information retrieval model
SoRank-A	Our model where only author social network is considered
SoRank-U	Our model where only user social network is considered

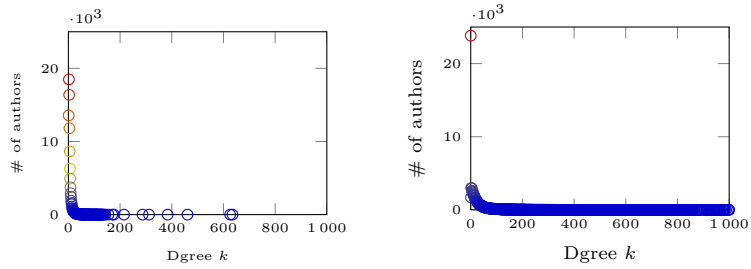
Table 5.5: Compared models for literature retrieval

5.6.2 Impact of the social network configuration on the retrieval effectiveness

In order to study the impact of network configurations on the retrieval effectiveness, we conduct here a comparative study on different author social networks topologies. We investigate the network extraction process as well as the type of integrated relationships.

The social network of CiteData corpus includes about 590000 co-authorships and 11 million citation relationships as previously shown in table 5.4. Accordingly, citations are 19 times more dense than co-authorships. Both networks follow however power-law distribution as show in figure 5.5.

Table 5.6 shows $P@20$ and MAP values obtained by ranking articles using only *PageRank* scores of respective authors. Rows determine the social relationships considered in the social network. Columns correspond to the implemented social network extraction method. The full social network includes all the authors in the corpus. A static *PageRank* score is assigned to each article regardless to the query. In the case of top authors network, the social network is build over the



(a) Co-authorship (b) Citation

Figure 5.5: Relationship distributions in the author social network

top 100 articles in the result set ranked by topical score as discussed in section 5.5.1.

	Full network		Top authors	
	P@20	*MAP*	*P*@20	*MAP*
Co-authorhips	0.230	0.108	0.314	0.127
Citation	0.246	0.112	0.316	0.139
Co-authorhips & Citation	**0.254**	**0.121**	**0.332**	**0.144**

Table 5.6: *P*@20 and *MAP* of PageRank with different network configurations

From table 5.6 we notice that for both full network and top author social network, citation relationships show better results rather than co-authorships. The combination of the two relationships leads to better *P*@20 and *MAP* values for both network sizes. As a result, citation relationships may express better the social importance of scientific researchers in the social network compared to co-authorships.

A considerable improvement is obtained in the case of top author network compared to the full social network of authors. In particular, a change of 30% of *P*@20 is ensured by *Co-authorship & Citation* configuration with top author network in comparison to full author network. This is explained by the fact that considering only authors from top results prevent from topical drift.

5.6.3 Evaluation of author's social importance

We compare our *SoRank-A* model with several baselines that rank scientific papers according to their scientific impact (*Cit*,*Exp-Cit*,*PR*) or the scientific impact of related authors (*h-index*,*PR-Cit*,*PR-Co*). The goal of these experiments

is to evaluate the social-network based measures to traditional scientific impact measures. As shown in the previous experiments in table 5.6, considering all the results for ranking is not helpful as re-ranking approaches may suffer from topic drift. In view of that, a ranking threshold is applied on the top 100 documents returned by the topical baseline, namely *BM25*. This setting is maintained for the rest of experiments.

	P@20		MAP	
Cit	0.262	44%	0.124	50%
Exp-Cit	0.246	54%	0.107	73%
PR	0.230	64%	0.107	73%
h-Index	0.262	44%	0.124	49%
PR-Cit	0.274	38%	0.112	65%
PR-CO	0.230	64%	0.108	72%
SoRank-A	**0.378**		**0.186**	

Table 5.7: Effectiveness of social networks measures and scientific impact measures

Comparison to scientific impact based metrics in table 5.7 shows that our *SoRank-A* ensures an improvement of 38% to 64% for *P*@20 measure and ensure an improvement of 49% to 73% for *MAP* measure.

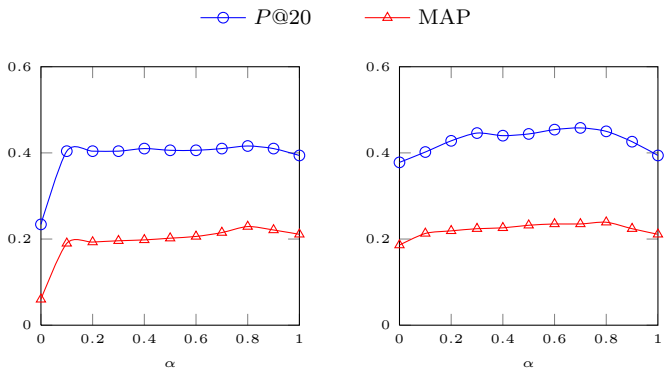
Among article-based impact metrics, namely citation index (*Cit*), expected citation (*Exp-Cit*) and article PageRank (*PR*), results show better performances by *Cit* model. In fact, these 3 models use citation links between articles as a basic feature to estimate the scientific impact of papers. Comparing, *PR* model to *PR-Cit* model which is a close related model that considers instead the citation network between authors, we notice that *PR-Cit* shows better results. Accordingly, author-based impact metrics may express better the relevance of articles rather than to article-based impact metrics.

In comparison to author-based impact metrics, typically social-network based metrics such as *PR-Co* and *PR-Cit* computed on author co-authorship network and author citation network, respectively. We note that *PR-Cit* shows better results as presented in the previous experiments. Our *SoRank-A* model presents however a considerable improvement for both evaluation measures. This is explained by the fact that our PageRank ranking algorithms consider in addition to the structure of the author social network their topical relevance as well. This expresses the expertise of the authors as discussed in section 5.4. Relevant authors may thus show a high social importance along with an expertise on the query topic.

5.6.4 Significance of author social network and user social network

In order to study the significance of both author social network and user social network and their ability to express the social relevance of scientific publications, we compare here *SoRank-A* and *SoRank-U* of our model that respectively consider the social network of authors and the social network of users in scholarly social bookmarking networks. Before discussing the performances of each model, we first proceed for α parameter tuning, presented in equation 5.9.

Linear combination parameter α helps to adjust importance of topical feature and social feature of our model. With $\alpha = 0$, documents are ranked using only the social feature, namely the social importance of either related author or users. Conversely, $\alpha = 1$ corresponds to the topical component of our model, namely $BM25$ model used in this case to retrieve candidate articles. Figure 5.6 presents the impact of α parameter for both measures $P@20$ and MAP.



(a) SoRank-U (b) SoRank-A

Figure 5.6: Tuning α parameter

Although α parameter do not have a considerable effect on *SoRank-U*, its impact on $P@20$ and MAP is more visible with *SoRank-A*. Both curves show peaks that exceed respective values obtained $\alpha = 0$ and $\alpha = 1$ that respectively correspond to the social component and the topical component of our model. We conclude that combining the topical relevance and the social importance of documents improves the retrieval process.

Best values of $P@20$ and MAP are achieved between $\alpha = 0.7$ and $\alpha = 0.8$. We conclude thus that the relevance of scientific papers depends primarily on the topical relevance. However, retrieval performances may be improved by considering the social importance of authors and users.

Figure 5.7 presents $P@20$ and MAP values obtained at $\alpha = 0.7$ and $\alpha = 0.8$ for *SoRank-U*, *SoRank-A* and *SoRank*. The *SoRank* model combines in fact the

social importance score of the two social networks of users and authors detailed in equation 5.14. For both evaluation measures, *SoRank-A* configuration shows better results than *SoRank-U*. The social network of authors expresses therefore the social importance of scientific publications compared to the social network of users. Meanwhile, this could be explained by the proportion of tagged documents in the dataset that represent only 22% of documents as shown previously in table 5.3.

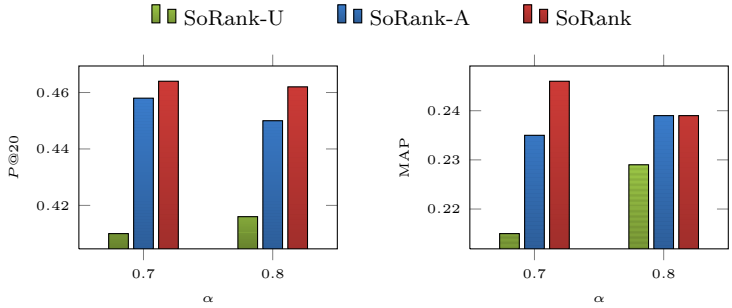


Figure 5.7: Effectiveness of author social network and user social network

Besides, results obtained by *SoRank* overpasses *SoRank-U* and *SoRank-A*. In particular, an improvement of 13% is obtained in comparison to *SoRank-U* for *P*@20 and an improvement of 4% is obtained in comparison to *SoRank-A* for *MAP* measure. This confirms our motivation for combining the two social networks of authors and users. In fact, the social importance of documents is defined by both social producing and social contexts. The two social contexts may mutually enhance each other.

5.6.5 Retrieval effectiveness

We compare in table 5.8 *P*@20 and *MAP* results obtained by our social model *SoRank* and the different baselines presented in table 5.5. Improvement of our *SoRank* is presented next to *P*@20 and *MAP* values. First, we note that our *SoRank* model overpasses all the baselines with significant results.

SoRank overpasses traditional information retrieval models with an improvement of 18% compared to *BM25* baseline and an improvement of 19% compared to *HiemLM* for *P*@30. This approves that our approach of combining the social relevance and the topical relevance enhances literature retrieval. We notice that the two baselines present the most competitive results to *SoRank*. This ca be explained by the fact that our model rely on the two models to compute the topical relevance of articles (*i.e.BM25*) and the expertise of authors (*i.e. HiemLM*).

		P@20			MAP	
BM25	0.394	18%	**	0.211	18%	**
HiemLM	0.388	19%	**	0.2064	20%	**
Cit	0.262	77%	***	0.124	100%	***
Expt-Cit	0.246	88%	***	0.107	132%	***
PR	0.23	101%	***	0.107	132%	***
h-Index	0.262	77%	***	0.124	100%	***
PR-CO	0.23	101%	***	0.108	130%	***
PR-Cit	0.274	69%	***	0.1121	121%	***
Kirsh	0.244	90%	***	0.1074	131%	***
SoRank	0.463		0.248			

Table 5.8: Comparing the retrieval effectiveness (*: $t.test < 0.05$; * * * : $t.test < 0.001$)

As we can see, *SoRank* presents better results compared to scientific impact measures that focuses on both articles (*CIT*,*Expt-CIT*,*PR*) and authors (*h-index*,*PR-CO*,*PR-Cit*). In comparison to results in table 5.7 where only the authors social importance is considered (*SoRank-A*), in other words $\alpha = 0$, we note a considerable improvement after α tuning. At this level, the relevance of articles is computed by combining topical and social features. This confirms, on the first hand, the interest of combining the two features for literature ranking as shown before with α tuning experiments. On the other hand, we conclude that the importance of authors and users may better express the social relevance in the context of bibliographic resources.

Comparing our model to the *Kirsch*'s model which is based on a social network approach, we note an improvement of 90%. In contrast of *Kirsch*'s model that uses co-authorship and combine relevance as the product of the social and topical relevant, our model uses a weighted co-authorship and citation network and combines scores with linear function. Thus, citation links, weighting network, and combining function have a considerable impact on the retrieval effectiveness.

5.7 Conclusion

We proposed in this chapter a social model for literature access that combines the topical relevance of scientific articles and the social importance of respective of authors and annotators. In particular, we propose to model authors with co-authorships and citation links. A weighting schema is attributed to

each social relationship in accordance to its type to express the shared interest, influence and knowledge transfer between authors. Users of scholarly bookmarking network are modeled with a social network where relationships are extracted from friendship and co-tagging activities. In order to evaluate the social importance of author, and similarity of bookmarking users, we propose a link analysis algorithm, named *"SoRank"*. Inspired by *PageRank*, our *SoRank* algorithm identifies authoritative actors in the social network while taking into account their expertise on the query topic.

We conduct a series of experiments on CiteData dataset. Experimental results show the interest of integrating citation link and co-authors in the social network of authors. In addition, we note the that social network build using authors of top articles with regard to the topic lead to more effective results with improvement of 30% compared to the entire social network of all authors. Our *SoRank* algorithm overpasses scientific impact metrics with an improvement of 44% compared to *Citation Index* and *h-index*. Compared to Citation *PageRank* and co-authorship *PageRank*, our model realize an improvement of 64%. Finally, we note that our model ensure a significant improvement of about 19% compared traditional information retrieval approaches and improvement of 90% compared to the Kirsch's which also propose the topical relevance of scientific paper with the social importance of authors.

Our global approach of combing the topical relevance and the social relevance could be applied for more application domains. In our paper (Ben Jabeur et al., 2011) we have extended this approach for microblogs retrieval.

In future work, we plan to extend social network of bibliographic resources with more entities involved in the scientific publication process. We plan also to integrate more social relevance features such as the publication date and the social distance between the person who submit the query and document's authors.

Chapter 6

Active microbloggers : Identifying influencers, leaders and discussers in microblogging networks

Introdcution

Microblogging services have emerged from a messaging application to news media (Phelan et al., 2009) and an open area for opinion expression (Jansen et al., 2009). With about 400 million[1] tweet published every day on Twitter, accessing to relevant messages that communicate fresh news and discuss interesting topics, becomes a challenging task. Users are overwhelmed by the huge quantity of useless, ambiguous, redundant and incredible posts. Ranking microblogs by chronological order is no longer appropriate for a better microblogging experience. Accordingly, and with respect to the following principle of microblogging networks where users are accessing to information through persons they follow, some researchs (Nagmoti et al., 2010; Das Sarma et al., 2010) have investigated the importance of microbloggers as a first step to access to interesting microblog posts. Tweet ranking problem is therefore viewed as microbloggers ranking task.

Ranking microbloggers consist of identifying important actors in the social network for a particular topic. This problem is defined by a ranking function $R(u_i, \mathcal{S}, G)$ that attributes a social importance score to each microblogger u_i in the social network G given a topic $\mathcal{S}$. This score evaluates the importance of a microblogger according to his position in the network. In the context

[1] https://blog.twitter.com/2013/celebrating-twitter7

of microblogs, previous works have addressed popularity (Duan et al., 2010), authority (Pal and Counts, 2011) and influence (Cha et al., 2010) as basic properties for important microbloggers. With the expansion of online communities, new types of microbloggers are attracting more attention thanks to their contributions and interactions in the social network. These microbloggers are called active and correspond to influencers as well as leaders and discussers.

Influencers are actors who are able to largely spread an information through the network. Leaders have the ability to motivate people and stimulate a community movement. Finally, discussers initiate valuable discussions around interesting topics. Besides influencers, leaders and discussers have not been yet investigated as key actors in microblogging networks. Meanwhile, some research on traditional blogs has addressed similar properties. In order to discover leaders, some works have focused on propagation patterns in the social network (Goyal et al., 2008). Discussers have been addressed in (Nakajima et al., 2006) as network agitators who simulate discussion in blog threads.

We are interested in this chapter in identifying active microbloggers and we propose a social network model that represents microbloggers using the social interactions between them such as following, retweeting and mentioning relationships. Moreover, we propose three different link analysis algorithms that highlight network influencers, leaders and discussers, respectively *InfRank*, *LeadRank* and *DiscussRank*. Our approach (Ben Jabeur et al., 2012b) is different from previous related work in at least two respects:

- We model the social network of microbloggers using a weighted multigraph that integrates followerships, retweets and mentions in the contrast of previous approaches using one or more binary social graphs (Kwak et al., 2010; Weng et al., 2010).
- We investigate influencers, leaders and discussers as key microbloggers unlike previous works focusing on popularity, authority and influence (Nagmoti et al., 2010; Kwak et al., 2010; Pal and Counts, 2011; Cha et al., 2010; Weng et al., 2010).

We introduce next the microblog information network then we describe the topology of the microblogger social network model. After that, we focus on active microblogger and we present ranking algorithms for identifying influencers, leaders and discussers in the social network. Finally we conduct a series of experiments on microblogging data in order to evaluate the performances of our ranking algorithms.

6.1 Microblogs information networks

The social information network of microblogs represents mutual interactions between actors and data. Network actors are represented by microbloggers. They play the role of information consumers and information producers. Network

data are mainly represented by microblogs. As shown in figure 6.1, there are three types of microblogs distinguished by functionality ad well as by Twitter user interface namely regular tweets, simply refered by *tweets*, replies and retweets. Replies are expendable to a conversion tree. Retweets are credited to the original author. Both replies and retweets inherit properties of regular tweets. In addition to tweets, network data include hashtags and Web resources. Although these entities are part of the tweet content, they are distinguished from the rest of tweet text by their syntax. Hashtags are preceded by # symbol. Web resources are represented by respective URLs.

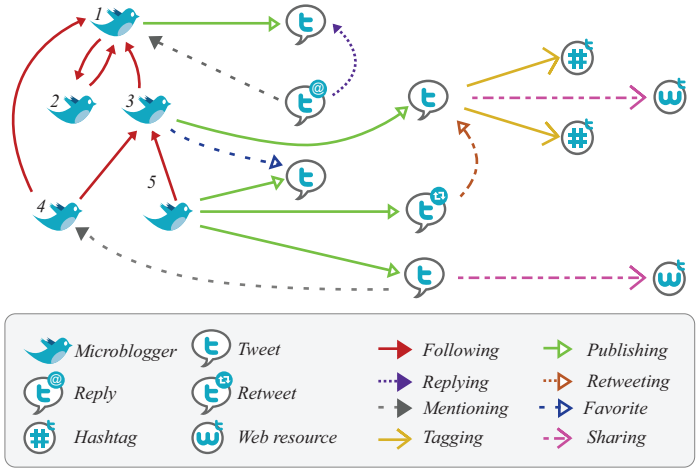


Figure 6.1: The social information network of Twitter

Microblog actors and data are connected with different types of relationships. We distinguish in figure 6.1 six types of relationships which are explicitly mentioned in Twitter. These relationships are classified into four categories according to the types of the involved entities:

Actor-to-Actor: represented by *followership* relationships.

Actor-to-Data: include *publishing* and *favorite* relationships.

Data-to-Data: include *retweeting*, *replying*, *tagging* and *sharing* relationships.

Data-to-Actor represented by *mentioning* relationships.

The above social relationships are explicitly defined in Twitter data. There are available through Twitter API[2]. For instance, *"followers/ids"*, *"friends/ids"* and *"favorites/list"* methods return the list of followers and followings and followers of a microblogger. API method *"statuses/show"* returns a structured description of the tweet including reference to retweeted status (*retweeted_status.id*),

[2]https://dev.twitter.com/

replied tweet (*in_reply_to_status_id*), mentions (*entities. user_mentions*), hashtags (*entities.hashtags*), urls (*entities.urls*), etc.

6.2 Microblogs social network

Based on the above social information network, we extract the social network that represents microbloggers and social relationships between them. In particular, the social network of microbloggers includes three types of relationships, namely, following, retweeting and mentioning relationships. Besides following relationships explicitly defined by microbloggers themselves, retweeting and mentioning relationships are inferred from respective links that connects tweets. We discuss in what follows the topology of social network of microbloggers and we define weight schemas on each type of the social relationship.

6.2.1 Network topology

We propose to represent the social network of microbloggers using a directed, labeled and weighted multigraph $G := (U, E, \Sigma_E, \ell_E, w)$ where:

- U is the set of microblogger nodes;
- $E = U \times U$ is the set of edges denoting relationships between microbloggers;
- $\Sigma_E = \{f, r, m\}$ is the alphabet of edge labels with f, r and m respectively following, retweeting and mentioning associations;
- $\ell_E : E \to \Sigma_E$ associates to each edge a label;
- $w : E \to \mathbb{R}$ associates to each edge a weight.

Unlike simple graphs, modeling social networks with multigraph allows to define multiple edges between network nodes. Microbloggers could be therefore connected with several social relationships simultaneously. Up to one edge from every relationship type may however defined between a couple of microbloggers.

$$\forall u_i, u_j \in U, r \in \Sigma_E \qquad |\{e : (u_i, u_j) \in E, \quad \ell_E(e) = r\}| \leq 1 \tag{6.1}$$

To avoid loops in the graph, typically when a microblogger mentions himself or reply to one of his previous tweets, edges pointing to the same node are discarded for the soical network.

$$\forall (u_i, u_j) \in E \qquad u_i \neq u_j \tag{6.2}$$

For each microblogger u_i and relationship type l, the set of node successors $\mathcal{O}(u_i, l)$ and the set of nodes predecessors $\mathcal{I}(u_i, r)$ are defined by:

- $\mathcal{O} : \mathcal{P}(U)$ associates to each microblogger $u_i \in U$ the set of successor nodes with connecting edges are labeled by $l \in \Sigma_E$,
- $\mathcal{I} : \mathcal{P}(U)$ associates to each microblogger $u_i \in U$ the set of predecessor nodes with connecting edges are labeled by $l \in \Sigma_E$.

6.2.2 Relationship weights

Let $T(u_i)$ be the set of tweets of microblogger u_i, $R^+(u_i)$ be the set of tweets retweeted by u_i and $M^+(u_i)$ be the set of tweets where u_i mentions another user. $R^+(u_i)$ and $M^+(u_i)$ are subsets of $T(u_i)$. Conversely, let $R^-(u_i)$ be the set of tweets of u_i retweeted by other users and $M^-(u_i)$ be the set of tweets where u_i has been mentioned. A weight is assigned to each following, retweeting and mentioning relationshipss as detailed next.

Following relationship: A followership edge $e : (u_i, u_j)$, with $\ell_E(e) = f$, is defined from a microblogger $u_i \in U$ to a microblogger $u_j \in U$ if the first microblogger follows the second one. This relationship shows the interest of u_i in the followed microblogger u_j. In order to quantify interest, we investigate reinforced followership between microbloggers, in other words, the number of intermediary nodes between them. This estimates the probability that the microblogger still receive the followed one's tweets even though respective followership is broken. The followership association is weighted as the proportion of intermediary nodes between microblogger over the social network. For convenience, we consider in addition to edge source u_i, direct intermediary nodes from first fellowship level only. Followership weight is defined by:

$$w_f(u_i, u_j) = \frac{|\mathcal{O}(u_i, f) \cap (\mathcal{I}(u_j, f) \cup \{u_i\})|}{|\mathcal{O}(u_i, f)|} \tag{6.3}$$

Retweeting relationship: A retweet edge $e : (u_i, u_j)$, with $\ell_E(e) = r$, is defined from a microblogger $u_i \in U$ to a microblogger $u_j \in U$ if there exists at least one tweet of u_j retweeted by u_i. Retweeting relationships reflect information diffusion and influence between microbloggers. This relationship would be as much reliable as microblogger u_i retweets tweets of microblogger u_j. Retweet exclusivity reflects however influence to respective microblogger compared to other retweeted ones. Accordingly, we propose to weight retweeting relationships with respect to overall tweets published by u_j. Retweeting weight is defined by:

$$w_r(u_i, u_j) = \frac{|T(u_j) \cap R^-(u_i)|}{|T(u_i)|} \tag{6.4}$$

Mentioning relationship: A mentioning edge $e : (u_i, u_j)$, with $\ell_E(e) = m$, is defined from a microblogger $u_i \in U$ to a microblogger $u_j \in U$ if there exists at least one tweet of u_i mentioning u_j. Mentioning relationships reflect information exchange and communication between microbloggers. A microblogger u_i would thus communicate as much information to a microblogger u_j as the second is mentioned in his tweets. On the other hand, mentioning a particular user rather than anyone else confirms focused communication between microbloggers. In view of that, we propose to weight mentioning relationships respectively to exclusive mentions between microbloggers. Weight on

mentioning edges is defined by:

$$w_m(u_i, u_j) = \frac{|M^+(u_i) \cap M^-(u_j)|}{|M^+(u_i)|} \tag{6.5}$$

Figure 6.2 illustrates a graph representation of the microblogger social network extracted from the information social network in figure 6.1. Edge weights are computed with respect to relationship types.

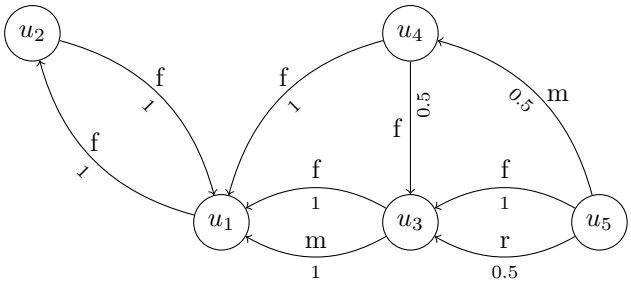


Figure 6.2: The social network of microbloggers.

6.3 Identifying active microbloggers

Based on the above social network model we define in this section the network properties of active microbloggers, typically influencers, leaders and discussers. A corresponding algorithm is proposed to identify each type of microbloggers.

6.3.1 Influencers

Influence in psychology is the process of changing of someone's emotion, opinion and behavior in accordance with the person who exercises influence (Raven, 1964). In the context of social networking environments, this change affects the networking activities of individuals. In fact, users my reproduce or discuss outstanding content from influential actors. Information carried by the original content is therefore diffused through the social network. According to this, network influencers are determined by their capacity to spread information through the social network.

Definition 10 *Influencers are active actors who have the ability to spread information and inspire other people in the network.*

The fact that information diffusion is an implicit indicator of its accuracy, influencers are therefore characterized with a high credibility. This gives more chance to their posts to interest other users in the network. In the context

of microblog networks, information spread is realized through retweet practice. However, the popularity of the microblogger contributes also to this process as it ensures a large visibility to the microblogger, and thus, gives more opportunity to his tweets to be diffused.

The popularity of a microblogger is defined by proportion of followed users over the social network. A popularity score is computed for each microblogger u_i as follows.

$$\mathcal{P}(u_i) = \frac{|\mathcal{I}(u_i, f)|}{|U|} \tag{6.6}$$

The influence of a microblogger is awarded by the number of retweets that he generates. However, this indicator quantifies influence at local level and do not study the entire social networks. On the other hand, microblogger influence is affirmed if he involves in retweets many microbloggers who, in their turn, have been retweeted frequently. This property remembers the principle of authority pages in the Web graph.

Inspired by the *PageRank* algorithm (Page et al., 1999), we propose to rank microbloggers by their mutual influence. In particular, we introduce a novel link analysis algorithm, called *InfRank*, that identifies authoritative microbloggers in the retweet network. Accordingly, good influencers are determined by microbloggers who have been retweeted by good influencers too. Microbloggers are assimilated here to Web pages while retweeting relationships are equivalent to hyperlinks in the Web graph. The detailed description of *Infrank* is given in algorithm 2.

Algorithm 2: InfRank

$k \longleftarrow 0$
foreach $u_i \in U$ **do** $Inf^k(u_i) = \mathcal{P}(u_i)$; `// initialization`
repeat
  $k \longleftarrow k + 1$
  foreach $u_i \in U$ **do**
    $Inf^k(u_i) = (1-d)\mathcal{P}(u_i) + d \times \sum_{u_j \in \mathcal{I}(u_i, r)} w_r(u_j, u_i) \frac{Inf^{k-1}(u_j)}{\mathcal{O}(u_j, r)}$
  end
  foreach $u_i \in U$ **do** $Inf^k(u_i) = \frac{Inf^k(u_i)}{\sum_{\forall u_j \in U} Inf^k(u_j)}$; `// normalization`
until *convergence* `// microblogger ranks never change`

InfRank defines microblogger's influence by two random walk probabilities. The first random walk probability is interpreted as the likelihood to randomly select a microblogger. This probability is computed based on the microblogger's

popularity as detailed in equation 6.6. The second random walk probability defines the likelihood of jumping from one microblogger to another by means of retweeting relationships. This probability is computed with respect of weights $w_r(u_i, u_j)$ defined on retweeting relationships as presented in equation 6.4. In particular, an influence score $Inf(u_k)$ is computed for each microblogger u_i using the next formula.

$$Inf^k(u_i) = (1-d)\mathcal{P}(u_i) + d \times \sum_{u_j \in \mathcal{I}(u_i,r)} w_r(u_j, u_i) \frac{Inf^{k-1}(u_j)}{\mathcal{O}(u_j, r)} \tag{6.7}$$

where k is the iteration index, d is dumping factor as defined *PageRank* algorithm that privileged either microblogger popularity or structural influence in social networks. $\mathcal{I}(u_i, r)$ and $\mathcal{O}(u_i, r)$ represent, respectively, retweeting predecessors and retweeting successors of microblogger u_i,.

In other words, *InfRank* computes microblogger's influence by propagating their popularity score through retweet edges. A microblogger accumulates so the popularity of other microbloggers who already have retweeted him. *InfRank* score is normalized at each iteration by the sum of *InfRank* scores of all microbloggers in the social network.

$$Inf^k(u_i) = \frac{Inf^k(u_i)}{\sum\limits_{u_j \in U} Inf^k(u_j)} \tag{6.8}$$

InfRank is computed iteratively until ranking convergence. This state is assumed whenever microblogger ranking remains the same for n consecutive iterations. For computation convenience, a minimum iteration number p and maximum iteration number q must be verified. Ranking convergence is achieved if the next condition is satisfied.

$$\forall u_i \in U, \quad Rank^k(u_i) = Rank^{k-1}(u_i) = \ldots = Rank^{k-n}(u_i) \tag{6.9}$$

where $k \in [p, q]$, $n < p$ and $Rank^k(u_i)$ is the rank of u_i at iteration k.

6.3.2 Leaders

Leadership is defined is psychology as a group activity where a person enlists the support from his community in order to accomplish a common goal (Chemers, 1997). In the context of microblogging networks, the intended goal is to shed light on a particular topic, usually an event or a cause, and engage a large number of people in the debate. In view of that initiative, leaders are determined by their capacity to create movements in the social network.

Definition 11 *Leaders are innovative actors, who take initiative, engage people and create movements in the a social network toward the introduced topic.*

In comparison to influencers, leaders have the ability to introduce new topics that evolve into public trends influencing a large number of microbloggers. Their tweets are likely to be retweeted and receive a lot of replies. In order to mobilize a large public, leaders must be characterized with high attraction which is represented be in fact by the size of their engaged community. This depends mainly on the number of followers as the case of influencers as well as the number of received retweets and replies. The attraction $\mathcal{A}(u_i)$ of a microblogger leader is defined by the proportion of following, retweeting and mentioning users over the social network.

$$\mathcal{A}(u_i) = \frac{|\mathcal{I}(u_i, f) \cup \mathcal{I}(u_i, r) \cup \mathcal{I}(u_i, m)|}{|U|} \tag{6.10}$$

The more a microblogger is followed, retweeted and mentioned by the others, the largest community he influences and mobilizes to reach a common goal. In accordance with the number of attracted users, a microblogger would acquire a high leadership potential. Meanwhile, leaders may engage other leaders too in order to increase his attraction. The more a microblogger is followed, retweeted and mentioned by the others, the largest community he influences and mobilizes to reach a common goal. In accordance with the number of attracted users, a microblogger would acquire a high leadership potential. Meanwhile, leaders may engage other leaders too in order to increase his attraction.

Based on the idea of the mutual leadership enhancement, we propose a *PageRank*-like algorithm, named *LeadRank*, that identifies enhanced leaders in the social network. Besides *InfRank* algorithm computed on the social network of retweets, *LeadRank* considers both retweeting and mentioning relationships. The detailed description of *LeadRank* is presented in algorithm 3.

Algorithm 3: LeadRank

$k \longleftarrow 0$

foreach $u_i \in U$ **do** $Ldr^k(u_i) = \mathcal{A}(u_i)$; `// initialization`

repeat

  $k \longleftarrow k + 1$

  foreach $u_i \in U$ **do**

    $$Ldr^k(u_i) = (1-d)\mathcal{A}(u_i) + d\times \left[\sum_{u_j \in \mathcal{I}(u_i, r)} w_r(u_j, u_i) \frac{Ldr^{k-1}(u_j)}{\mathcal{O}(u_j, r)} \times \sum_{u_j \in \mathcal{I}(u_i, m)} w_m(u_j, u_i) \frac{Ldr^{k-1}(u_j)}{\mathcal{O}(u_j, m)} \right]$$

  end

  foreach $u_i \in U$ **do** $Ldr^k(u_i) = \frac{Ldr^k(u_i)}{\sum_{\forall u_j \in U} Ldr^k(u_j)}$; `// normalization`

until *convergence* `// microblogger ranks never change`

LeadRank represents the leaderships of a micrcroblogger using two random walk probabilities. The first random walk probability is interpreted as the likelihood to randomly select a microblogger. This probability is computed based on the microblogger's attraction $\mathcal{A}(u_i)$ as detailed in equation 6.10. The second random walk probability defines the likelihood of jumping from one microblogger to another by means of retweeting and mentioning relationships. This probability is computed with respect of weights $w_r(u_i, u_j)$ and $w_m(u_i, u_j)$ defined on retweeting and mentioning relationships as presented respectively in equations 6.4 and 6.5. *LeadRank* computes a leadership score $Ldr(u_k)$ for each microblogger u_i as follows.

$$\begin{aligned} Ldr^k(u_i) = (1-d)\mathcal{A}(u_i) + d \\ \times \sum_{u_j \in \mathcal{I}(u_i,r)} w_r(u_j,u_i) \frac{Ldr^{k-1}(u_j)}{\mathcal{O}(u_j,r)} \\ \times \sum_{u_j \in \mathcal{I}(u_i,m)} w_m(u_j,u_i) \frac{Ldr^{k-1}(u_j)}{\mathcal{O}(u_j,m)} \end{aligned} \tag{6.11}$$

where k is the iteration index. d is dumping factor as defined *PageRank* algorithm. $\mathcal{I}(u_i, r)$, $\mathcal{O}(u_i, r)$, $\mathcal{I}(u_i, m)$ and $\mathcal{O}(u_i, m)$ represent, respectively, retweeting predecessors, retweeting successors, mentioning predecessors and mentioning successors.

LeadRank estimates the leadership of microbloggers by propagating their attraction weights through incoming retweets and mentions. In particular, the leadership score is computed as the product of two sums. The first sum accumulates the leadership scores of retweeting microbloggers. It estimates the influence of the microblogger. The second sum is conducted over the leadership scores of mentioning users. It highlights engagement over the mobilized community. Multiplying the two sums ensures both properties of leaders, namely influence and community mobilization. Replacing the product by a simple sum ends either to influencers or authoritative microbloggers in mentioning social networks.

Similarity to *InfRank*, the leadership score $Ldr^k(u_i)$ is normalized by division by the sum of all leadership scores. Once again, algorithm convergence is assumed when node ranking remains stable for n consecutive iterations as presented in equation 6.9

6.3.3 Discussers

Discussers are in boarder context are persons who take up in conversation or in a debate. They conduct a close examination of a subject with interchange of opinions. This is insured in microblog networks by means of replies and mentions. Besides the messaging purpose, microblog discussers have the ability to initiate interesting conversations.

Definition 12 *Discussers are interactive actors who initiate valuable conversations around an outstanding content in a social network.*

Tweet published by a discusser may interest other users in the microblogging network since it enriches the original post with further details, fresh updates and even rectification for wrong information. As they track valuable tweets to discuss, following a discusser may be helpful to get acknowledged with the interesting tweets and debates. The importance of a discusser is defined by the number his interlocutors, namely motioned and motioning microbloggers. This reflects the conversational interactions he has established. The conversational interaction of $\mathcal{C}(u_i)$ a microblogger is computed so as the proportion of motioned and motioning users of the social network.

$$\mathcal{C}(u_i) = \frac{|\mathcal{I}(u_i, m) \cup \mathcal{O}(u_i, m)|}{|U|} \tag{6.12}$$

Microbloggers already in interaction with many interlocutors are potential candidates of network discussers. The reciprocal interaction between discussers increases their chance to be good discussers. Accordingly, we propose a link analysis algorithm, called *DiscussRank*, that highlights mutually connected discussers in social network. *DiscussRank* is a modified version of *PageRank* algorithm computed on the mentioning network which takes into account the incoming and outgoing edges. The detailed description of *DiscussRank* is given in algorithm 4.

Algorithm 4: DiscussRank

$k \longleftarrow 0$

foreach $u_i \in U$ **do** $Desc^k(u_i) = \mathcal{C}(u_i)$; `// initialization`

repeat

  $k \longleftarrow k + 1$

  foreach $u_i \in U$ **do**

    $Desc^k(u_i) = (1-d)\mathcal{C}(u_i) + d\times$

    $\left[\sum_{u_j \in \mathcal{I}(u_i,m)} w_m(u_j, u_i) \frac{Desc^{k-1}(u_j)}{\mathcal{O}(u_j,m)} \times \sum_{u_j \in \mathcal{O}(u_i,m)} w_m(u_i, u_j) \frac{Desc^{k-1}(u_j)}{\mathcal{I}(u_j,m)}\right]$

  end

  foreach $u_i \in U$ **do** $Desc^k(u_i) = \frac{Desc^k(u_i)}{\sum_{\forall u_j \in U} Desc^k(u_j)}$; `// normalization`

until *convergence* `// microblogger ranks never change`

DiscussRank estimates the importance of discussers using two random walk probabilities. The first random walk probability is interpreted as the likelihood to randomly select a microblogger. This probability is computed based on the conversational interaction $\mathcal{C}(u_i)$ defined in equation 6.12. The second random

walk probability defines the likelihood of jumping from one microblogger to another by means of mentioning relationships. This probability is computed with respect of weights $w_m(u_i, u_j)$ defined on retweeting relationships as presented in equation 6.5. A discusser score $Desc(u_i)$ is computed iteratively for each microblogger u_i using the next formula.

$$\begin{aligned} Desc^k(u_i) = (1-d)\mathcal{C}(u_i) + d \\ \times \sum_{u_j \in \mathcal{I}(u_i,m)} w_m(u_j, u_i) \frac{Desc^{k-1}(u_j)}{\mathcal{O}(u_j, m)} \\ \times \sum_{u_j \in \mathcal{O}(u_i,m)} w_m(u_i, u_j) \frac{Desc^{k-1}(u_j)}{\mathcal{I}(u_j, m)} \end{aligned} \tag{6.13}$$

Where k is the iteration index, d is the random walk parameter. $\mathcal{I}(u_i, m)$ and $\mathcal{O}(u_i, m)$ represent, respectively, mentioning predecessors and mentioning processors of microblogger u_i,.

DiscussRank algorithm propagates iteratively discusser scores via incoming and outgoing mentioning edges. The product of the two sums of mentioned and mentioning microblogger scores is maximized if the discusser is, at the same time, mention and been mentioned by many microbloggers. Accordingly, *DiscussRank* identifies simultaneous authorities and hubs nodes in the social network. Similarity to previous algorithms, discusser score $Desc^k(u_i)$ is normalized by division by the sum of all leadership scores. Once again, algorithm convergence is assumed when node ranking remains stable for n consecutive iterations as presented in equation 6.9.

6.4 Experimental evaluation

We conduct a series of experiments on microblog data in order to study the performances of our three ranking algorithms: *InfRank*, *LeadRank* and *DiscussRank*. The main goals of these experiments are to examine the ranking process of these algorithms and evaluate their effectiveness for indentifying interesting microbloggers in the social network. We present in what follows the evaluation protocol and then we discuss results and findings.

6.4.1 Experimental setup

In order to indentify active users in microblog networks, we carry out a user study on a tweet corpus from TREC 2011 Microblog dataset. We build for this aim a topic dataset from major events in the corpus and we involve regular twitter users in order to rate the ranking results of our proposed algorithms and compared baselines.

6.4.1.1 Tweet corpus.

We used in these experiments the TREC 2011 Microblog dataset *Tweets2011*. This corpus is crawled over 16 days from January, 23^{rd} to February, 8^{th}, 2011. Further details about *Tweets2011* corpus and TREC 2011 Microblog track are discussed in section 4.5. *Tweets2011* corpus includes about 16 million tweets and over 5 million microbloggers with approximately 3 tweets per user. Table 6.1 presents general statistics of microblog entities in the corpus.

Tweets	16 141 812	Microbloggers	5 356 432
Retweets	1 128 179	Hashtags	2 466 654
Mentions	7 193 656	URLs	2 769 955

Table 6.1: Tweets2011 statistics

6.4.1.2 Topic dataset.

We are interested in these experiments in microblogger search task. This task aims to identify key microbloggers with regard to a particular topic. Unlike TREC Microblog real-time search task that ranks relevant tweets to a user query, microblogger search task deals with microbloggers instead of tweets. To perform this task, we defined 3 topics from main events that inspired microbloggers during the corpus period. We retrieve from the corpus all tweets that contain at least on term of the query topic. We build afterward the microblogger social network from extracted tweets. Table 6.2 lists proposed topics as well as statistics about tweets, microbloggers, followership, retweeting and mentioning relationships. We note that *tweets2011* corpus does not include followership information. We crawled this data using Twitter API.

#	Topic	Tweets	Users	Follow.	Ret.	Men.
1	NFL Super Bowl	55 225	52 082	41 695	951	23 674
2	Egypt's Tahrir Square protests	53 047	36 571	154 628	27 712	12 976
3	State of the Union address	21 986	20 068	15 673	541	221
Mean		43 419	36 240	70 665	9 735	12 290

Table 6.2: Topic dataset statistics

6.4.1.3 Baselines.

We compare our proposed algorithms to the next 3 baselines:

– *followers* baseline ranks microbloggers by respective number of followers. This indicator reflects the polarity of the microblogger in the social network.

– *f-pagerank* baseline ranks microbloggers according to their authority. It computes a *PageRank* score on followership network (Kwak et al., 2010).

– *r-pagerank* baseline ranks microbloggers according to their influence. It computes a *PageRank* score on the retweeting social network (Duan et al., 2010).

6.4.1.4 Results assessment.

Inspired by the evaluation protocol proposed by Pal and Counts (2011), we asked 2 regular Twitter users (a_1 and a_2) to rate the interestingness of 360 microbloggers from the top 20 results returned by each algorithm and baseline. In addition to classic topical relevance, microblogger interestingness takes into account microblogging practices (*e.g.*, appropriate use of hashtags, mentions, *etc*), tweet quality (*e.g.*, language, style, attitude *etc*) and the microblog profile (*e.g.*, description, prestige, *etc*). A rate between 0 and 2 is assigned to each microblogger in accordance to his interestingness: (0 *Not interesting*; 1 *Minimally interesting*; 2 *Highly interesting*).

In order to study the impact of the social context on microblogger rating, typically this evaluation is conducted in two steps.

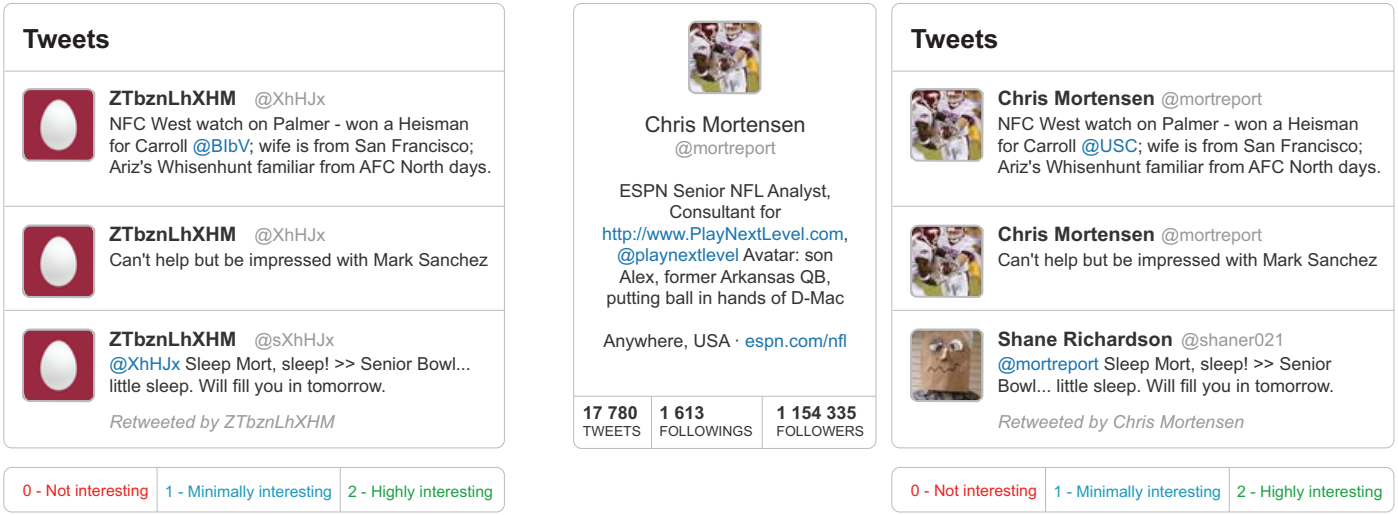
– *Anonymous evaluation AI:* The interestingness of microbloggers is evaluated in this step based only on their tweets. Microblogger information including description, avatar and the number of followers are hidden. Furthermore, profile name and tweets mentions are passed to encryption.

– *Non-anonymous evaluation ¬AI:* The interestingness of microbloggers is evaluated in this step based on both tweet content and microblogger profile. In particular, microblogger name and tweet mentions are revealed. In addition, microblogger description, avatar and the number of followers are displayed in addition to tweets. In the case of retweets, we display the original author's name.

More realistic experience, annotation interface was designed similarly to Twitter user interface. Figure 6.3 shows a screen capture of user rating for both non-anonymous and anonymous evaluation.

Table 6.3 shows inter-annotator and inter-evaluation agreements measured by Cohen's κ coefficient. A moderate inter-annotator agreement of 0.573 are obtained along the two evaluations processes. A slightly higher agreement is generated in the case of non-anonymous evaluation. On the other hand, a strong agreement is shown between the two evaluation settings with a value of Cohen's κ equal to 0.709. This demonstrates the importance of topical relevance for microblogger interestingness evaluation.

For the rest of experiments, a summarized interestingness rate is attributed



(a) Anonymous evaluation

(b) Non-Anonymous evaluation

Figure 6.3: Anonymous evaluation versus Non-Anonymous evaluation

to each microblogger as the rounded mean rate given by the 2 annotators $mean(a_1, a_2)$.

(a) Inter-annotator

	a_1 vs a_2
AI	0.569
¬AI	0.577
AI & ¬AI	0.573

(b) Inter-evaluation

	AI vs ¬AI
a_1	0.730
a_2	0.685
$mean(a_1, a_2)$	0.709

Table 6.3: Inter-annotator & inter-evaluation agreement (Cohen's κ coefficient)

6.4.2 Ranking correlation

In order to measure the degree similarity between ranking algorithms, we present in table 6.4 the Kendall's τ coefficient for each pair of ranking models. Extreme values of correlation coefficient show either total disagreement as $\tau = -1$ or perfect agreement between ranking models with $\tau = 1$. Nevertheless, two ranking models are assumed independent if respective correlation coefficient is equal to 0. This is shown for example by *followers* rankings where τ value is always near to 0. This baseline is therefore independent from all the compared models.

		R1	R2	R3	R4	R5	R6
R1	*followers*	1.00	0.08	−0.02	0.07	0.08	−0.02
R2	*f-pagerank*	0.08	1.00	0.49	**0.85**	**0.84**	0.45
R3	*r-pagerank*	−0.02	0.49	1.00	0.57	0.61	0.53
R4	InfRank	0.07	**0.85**	0.57	1.00	**0.89**	0.41
R5	LeadRank	0.08	**0.84**	0.61	**0.89**	1.00	0.41
R6	DiscussRank	−0.02	0.45	0.53	0.41	0.41	1.00

Table 6.4: Rank correlation measured by Kendall's τ coefficient

Even though *f-pagerank*, *r-pagerank* and DiscussRank use different types of networks, respectively, following network, retweet network and mention network, they show, surprisingly, a fair agreement with correlation coefficient is close to 0.5. An inside look on ranking correlation over the top results in figure 6.4 shows that the three algorithms present either low correlated or different rankings. In fact, correlation coefficient value is explained due to the large number of microbloggers in rankings tail that were ranked similarly since they are involved in no social relationship.

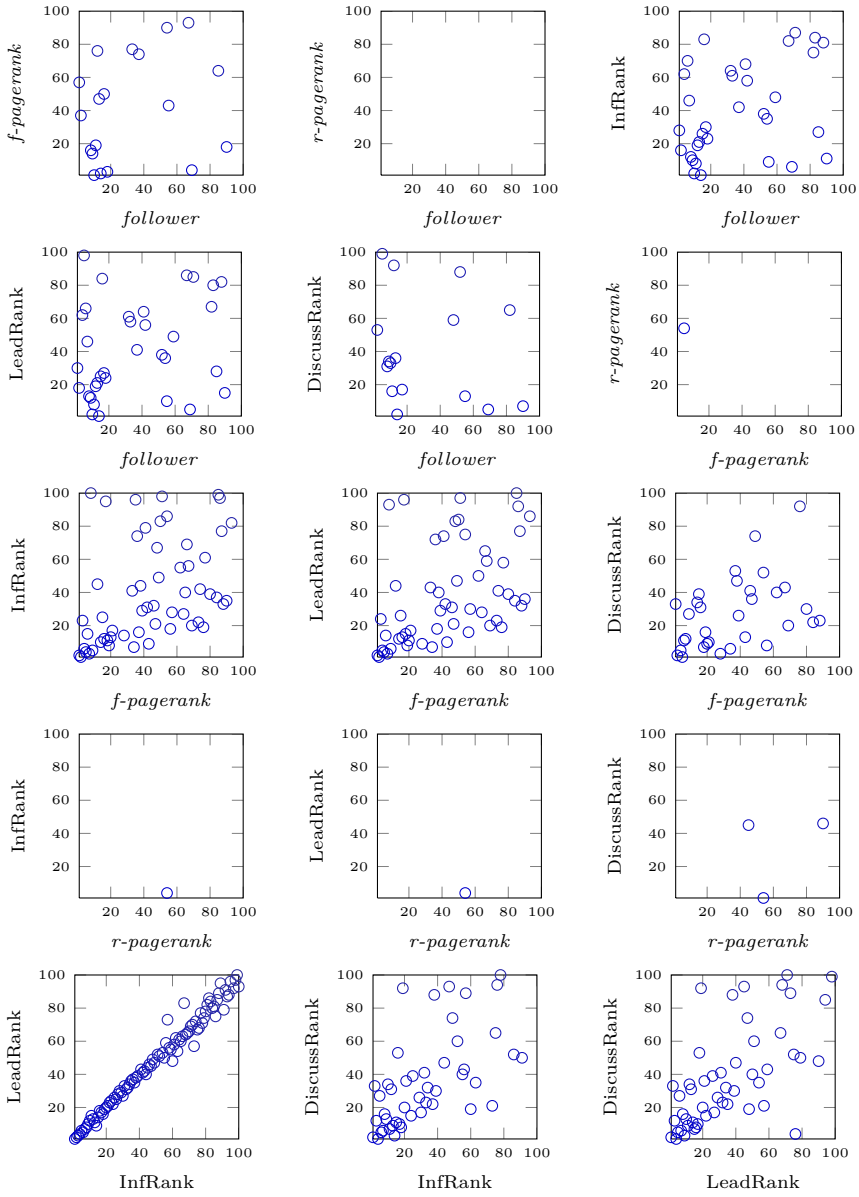


Figure 6.4: Ranking correlation for topic 2

A high agreement is shownon the first hand, between *f-pagerank* and InfRank, and on the other hand between *f-pagerank* and LeadRank. A notable correlation is also shown over the top 100 ranking as illustrated in figure 6.4. Although InfRank and LeadRank do not exploit followership links, they take into account the number of followers as an indicator of influence and leadership. It allows to accomplish comparable performance to *f-pagerank* while keeping independent rankings to *followers* model. This is advantageous once following social network is too costly in terms of crawling and exploring.

Comparing the three proposed algorithms, InfRank and LeadRank show a considerable agreement with the Kendall's τ coefficient is equal to 0.89. This is explained by the fact that the two algorithms use similar features namely retweeting relationships. This result confirms in fact the motivation behind the two ranking models as both of them address the influence of microbloggers. DiscussRank which addresses however another property of active microbloggers related to their conversational activities, shows fair agreement to the two algorithms with a correlation coefficient of 0.41.

6.4.3 Active microblogger precision

We study at this level the performances of our 3 ranking algorithms, in particular the ability to return interesting microbloggers in the top 5, 10 and 20 results. Table 6.5 presents precisions $P@5$, $P@10$ and $P@20$ of microblogger interestingness for both anonymous evaluation AI and non-anonymous ¬AI evaluation. We note that *LeadRank* algorithm shows higher results for all the evaluation settings and precisions. For $P@10$ and $P@20$ which are more significant precision is near to 0.6. *InfRank* shows close performances with $P@10 = 0.53$ and $P@20 = 0.47$ for AI and $P@10 = 0.43$ and $P@20 = 0.45$ for ¬AI. These 2 algorithms seem therefore more accurate. Accordingly, influence based ranking is more relevant for active microblogger rankings.

	AI			¬AI		
	$P@5$	$P@10$	$P@20$	$P@5$	$P@10$	$P@20$
InfRank	0.47	0.53	0.47	0.47	0.53	0.45
LeadRank	**0.73**	**0.60**	**0.57**	**0.80**	**0.63**	**0.57**
DiscussRank	0.33	0.43	0.40	0.33	0.47	0.38

Table 6.5: Precision of active microbloggers algorithms

Table 6.6 lists the top 10 microbloggers returned by each ranking algorithm. Unique microbloggers from each ranking are written in bold. We note that some news providers, blogs, and celebrities are highly ranked such as *@CNN* (news media), *@nfl* (sport league) and *@rickyrozay* (artist). We conclude so

that real world popularity may enhance active microblogger position in the social network. Furthermore, we note that the three algorithms present some common microbloggers for the same topic such as *@monaeltahawy* (journalist) in topic 2 or for many topics such as *@Reuters* (news agency) in topic 2 and 3. These microbloggers are characterized thus by an important social activity allowing them to have a strategic position in the social network.

Topic	Rank	InfRank	LeadRank	DiscussRank
1	1	***@Karo_Dita***	*@KhloeKardashian*	*@nfl*
	2	***@vanessavillazon***	*@nfl*	***@Deeener***
	3	*@KhloeKardashian*	*@espn*	***@BeliebersGirl***
	4	*@CNN*	*@CNN*	***@KobaMyColored***
	5	*@espn*	*@rickyrozay*	***@thelovatobrasil***
	6	*@nfl*	*@mortreport*	***@FC_MyLife_LS***
	7	*@rickyrozay*	*@LMFAO*	***@JhenneVieira***
	8	*@mortreport*	*@KREAYSHAWN*	***@eternadrenalina***
	9	*@LMFAO*	***@iamwill***	***@amand4___***
	10	*@KREAYSHAWN*	***@nflnetwork***	***@bep***
2	1	*@AJEnglish*	*@AJEnglish*	*@monaeltahawy*
	2	*@Reuters*	*@Reuters*	*@AJEnglish*
	3	*@BreakingNews*	*@BreakingNews*	*@AymanM*
	4	*@monaeltahawy*	*@monaeltahawy*	***@speak2tweet***
	5	*@nytimes*	*@SultanAlQassemi*	*@SultanAlQassemi*
	6	*@SultanAlQassemi*	*@nytimes*	*@bencnn*
	7	*@bencnn*	*@bencnn*	***@alaa***
	8	*@NickKristof*	*@NickKristof*	***@sharifkouddous***
	9	*@AJELive*	*@AymanM*	***@CNN***
	10	***@BBCWorld***	*@AJELive*	***@Dima_Khatib***
3	1	***@Mumiangel***	*@Reuters*	***@JenEngland***
	2	***@qyrrrAE***	*@TheEconomist*	***@ilikesleep***
	3	***@egothai***	*@BBCWorld*	***@SultanAlQassemi***
	4	*@Reuters*	***@AJELive***	*@BBCWorld*
	5	@PaulMBaker	***@politico***	*@StateDept*
	6	*@TheEconomist*	***@CBSNews***	***@steffensmark***
	7	***@SwapnilTalekar***	*@StateDept*	*@ABC*
	8	***@nickymatonak***	*@ABC*	***@Dima_Khatib***
	9	***@rocaral***	***@UN***	***@Elicoopter_mid***
	10	*@sarahmeeks24*	*@SultanAlQassemi*	*@HBCUDigest*

Table 6.6: Top 3 microbloggers returned by InfRank, LeadRank and Discuss-Rank

Regardless of their popularity, some microbloggers are ranked better than well known microbloggers. For instance, InfRank ranks *@PaulMBaker* (researcher, $1K$ followers) before *@TheEconomist* (magazine, $3.3G$ followers) in topic 3. This microblogger may however publish tweets that report own point of view that subsequently influence other users. On the other hand, news media microbloggers dominate LeadRank rankings typically for topics 2 and 3 which are news-based topics. Obviously, major news providers such as *@Reuters* and *@AJEnglish* (Middle East news channel) may act as leaders for new coverage about the two events.

Regarding topic 2 in relation to pro-democracy movement, DiscussRank high-

lights some notable microbloggers namely *@speak2tweet*, a communications service developed by Google to help people reporting news as Internet was shut down by Egyptian government. Another microblogger identified by DiscussRank algorithm is *@alaa*, Egyptian activist and one of the important actors during this massive movement. Both of microbloggers are characterized by intensive communication activities with regards to this topic.

6.4.4 Comparison with related models

In order to evaluate the effectiveness of our algorithms, we compare in table 6.7 Normalized Discounted Cumulative Gain (NDCG) values for baselines and algorithms rankings and this for anonymous evaluation AI non-anonymous evaluation ¬AI. NDCG measure evaluates in fact ability of a model to rank relevant results in accurate order, namely by decreasing relevance rates.

First, we notice that AI rating shows slightly different NDCG@10 and NDCG@20 values to ¬AI ratings . With the social context is revealed, annotators evaluate stricter interestingness of microbloggers by considering profile data and social interactions. In the case of AI evaluation, *LeadRank* presents highest NDCG@10 and NDCG@20 values with $NDCG@10 = 0.15$ and $NDCG@20 = 0.24$. Close results are shown by *followers* model with $NDCG@10 = 0.14$ and $NDCG@20 = 0.19$. Other models are, however, less effective.

	AI			¬AI		
	$NDCG@5$	$NDCG@10$	$NDCG@20$	$NDCG@5$	$NDCG@10$	$NDCG@20$
followers	0.10	0.14	0.19	0.10	0.14	0.19
f-pagerank	0.05	0.06	0.08	0.06	0.07	0.10
r-pagerank	0.03	0.04	0.08	0.04	0.05	0.08
InfRank	0.05	0.10	0.15	0.06	0.13	0.18
LeadRank	**0.11**	**0.15**	**0.24**	**0.14**	**0.18**	**0.27**
DiscussRank	0.06	0.11	0.16	0.00	0.04	0.11

Table 6.7: Comparison of baseline effectiveness for AI and ¬AI evaluations

Considering ¬AI evaluation which is the more significant one in these experiments, we note that InfRank and LeadRank algorithms present higher values. We conclude that influence is a primordial property of active microbloggers. Similar performances are shown by R-PageRank algorithm which investigates also microblogger influence with different interpretations.

Conclusion

We present in this chapter a microblog social network model represent microbloggers with different social interaction including followership, retweets and mentions.

We proposed in this chapter a weighted social network model that represents microbloggers and their mutual social interactions. Furthermore, we proposed microblogging specific link-analysis algorithms that identify influencers, leaders and discussers in the network. In particular, proposed *InfRank*, *LeadRank* and *DiscussRank* algorithms compute a social score for each microblogger by propagating weights through social network. Experiments on TREC 2011 Microblogs show that *LeadRank* algorithm overpasses others algorithms and baselines.

In future work, we plan to evaluate our algorithms with additional approaches in the literature. We also plan to integrate the proposed algorithms into a real-time content discovering system that focuses on active microbloggers instead of time-costly approaches analyzing all microblogs.

Chapter 7

Featured tweet search: Modeling time and social influence for microblog retrieval

Introduction

Microblogs are popular networking services that enable users to broadcast an information. Unlike news headlines which is generated by mass media, microblogs address general topics that interest a large public as well as small communities and close social networks. In addition, microblogs enrich reported news with valuable information. For instance, some particular events are covered in real-time with instant updates and live photos from the event site. Moreover, microblogs identify the exact source of information (author) and describe its publishing context (time, geolocalisation, application, device, etc). Finally, microblogs extended the informative purpose of message broadcasting and enable people to express their opinion about real world events.

With the variety of supported features, microblogging services emerge as a promising tool to get acquainted with the latest news. However, seeking for information over microblogging spaces becomes a challenging task due the increasing amount of published information. In the case of Twitter microblogging service, which is the focus of this work, about 400 million[1] messages (called *"tweets"*) are published every day. A part of these tweets are useless, ambiguous, redundant or incredible (Sankaranarayanan et al., 2009). A new infor-

[1] https://blog.twitter.com/2013/celebrating-twitter7

mation retrieval task is therefore created. Its main purpose is to search for real-time information and to rank recent tweets. TREC 2011 Microblog track (Ounis et al., 2011) defines tweet search as a real-time adhoc task where the users are interested in most recent and relevant information. In the spite of Web search, tweet search aims to find temporally relevant information, monitor content and follow current events and people activities (Teevan et al., 2011).

Prior works addressing tweet search integrate a variety of textual features, microblogging features and social network features (Nagmoti et al., 2010; Duan et al., 2010). These works consider that tweet relevance depends, on the one hand, from the importance of corresponding authors in the social network and, on the other hand, from the content quality such as URLs, mentions and hashtags. We investigate in this chapter different motivations behind tweet search, namely topical, temporal and social motivations. We propose an integrated Bayesian network model that considers:

- the number of query terms in the tweet as an indicator of topical overlap between the query and the tweet;
- the social importance of the related microblogger as an indicator of tweet credibility;
- the tweeting features such as hashtags, mentions, and URLs;
- the topic activity periods which corresponds to the joint events in the real world.

Unlike related work, our model is characterized by the following features:

- Tweet relevance estimation is addressed using a Bayesian network model that integrates all used features. Previous work uses clustering-based approaches or learning to rank methods to combine separated features (Sankaranarayanan et al., 2009; Duan et al., 2010; Nagmoti et al., 2010).
- Microbloggers are represented with several relationships including followerships, retweets and mentions in the contrast of the followers' social network used in (Kwak et al., 2010; Weng et al., 2010). Moreover, we consider only the sub-network generated by retrieved tweets avoiding so the dominance of some celebrities if the entire social network is considered (Duan et al., 2010).
- The time magnitude of a tweet is estimated from each term's occurrence in the temporal neighborhood in the contrast of work in (Grinev et al., 2009) analyzing all tweets to locate activity burst periods of a specific topic.

In particular, we propose two topologies of Bayesian network models for tweet search. The first model is based on inference networks (Ben Jabeur et al., 2012e). The second model is based on belief networks (Ben Jabeur et al., 2012c,d). The motivation behind the use of Bayesian network is that this family of models supports the dependency between the integrated features. Tweet search is a particular information retrieval task driven by a variety of topical, social and temporal motivations that may mutually dependent. Bayesian network models ensure the retrieval process even though some data is unavailable such as a protected microblogger profile.

We define in what follows common definitions and notations by the two Bayesian network models for tweet search. Afterward, we focus on each tweet Bayesian network model for tweet search and we detail network topology as well as the query evaluation process. Finally we conduct a series of experiments using TREC Microblog Track in order to validate our proposed models.

7.1 Definitions and notations

Term: let K be the set of terms. Each term $k_i \in K$ is associated to a random variable $k_i \in \{0,1\}$. The event of *"observing term k_i"* is noted $k_i = 1$ or shortly k_i. The complement event that *" term k_i is not observed"*, is noted $k_i = 0$ or shortly $\bar{k_i}$. We notice that the same notation "k_i" is used to represent term k_i as well as respective random variable k_i.

Let p be the number of index terms. It exists 2^p possible combinations of terms, called *term configurations*. A configuration $\vec{k}$, may represent a tweet or a query. An index of 2 terms (k_1, k_2) presents for instance $2^2 = 4$ possible configurations represented by the next set:

$$\mathscr{C} = \left\{(k_1, k_2), (k_1, \bar{k_2}), (\bar{k_1}, k_2), (\bar{k_1}, \bar{k_2})\right\} \tag{7.1}$$

Let $\vec{k}$ be a term configuration, we define $c(\vec{k})$ and $on(k_i, \vec{k})$ as follows:

- $c(\vec{k})$ associates to a configuration k_i, the set of positively instantiated terms.
- $on(k_i, \vec{k})$ associates to a configuration k_i, the value of corresponding random variable k_i. $on(k_i, \vec{k}) = 1$ if term k_i is positively instantiated in $\vec{k}$. Conversely, $on(k_i, \vec{k}) = 0$ if term k_i is not instantiated in $\vec{k}$.

Considering for instance the configuration $\vec{k} = (k_1, \bar{k_2}, k_3)$. The set of positively instantiated terms is defined by $c(\vec{k}) = \{k_1, k_3\}$. On the other hand, the values of respective random variables are defined by $on(k_1, \vec{k}) = 1$, $on(k_2, \vec{k}) = 0$ and $on(k_3, \vec{k}) = 1$.

Tweet: Let T be the set of tweets. Each tweet $t_j \in T$ is associated to a random variable $t_j \in \{0,1\}$. The event $t_j = 1$ of *"observing tweet t_j"* is noted t_j. The complement event $t_j = 0$ is noted $\bar{t_j}$. A tweet t_j is represented as a term configuration $t_j = (k_1, ..., k_i, ..., k_n)$ with k_i is a random variable indicating if either term k_i is present in the tweet or not. Similarly to term notation, "'t_j" is used to represent tweet t_j as well as respective random variable t_j.

tf_{k_i,t_j} indicates the frequency of term k_i in tweet t_j. If k_i is present in t_j, $tf_{k_i,t_j} > 0$. Otherwise, tf_{k_i,t_j} is set to 0. Tweet length tl_{t_j} is defined by the number of terms in tweet t_j. In other words, tl_{t_j} corresponds to the sum of frequencies of tweet terms $tl_{t_j} = \sum_{\forall k_i} tf_{k_i,t_j}$.

Query: In the same way as tweets, a query is represented as a term configuration $q = (k_1, ..., k_i, ..., k_n)$ with k_i a random variable indicating if either term k_i is present in the query. Let Q be a set of user queries. A query $q \in Q$ is associated to a random variable $q \in \{0, 1\}$. The event $q = 1$ of *"observing query q"* is noted q. The complement event $q = 0$ is noted $\bar{q}$.

Microblogger: Let U be the set of microbloggers. Each microblogger $u_f \in U$ is represented by a random variable $u_f \in \{0, 1\}$. The event $u_f = 1$, shortly written as u_f, denotes *"microblogger u_f is observed"*. $u_f = 0$, noted $\bar{u}_f$, denotes *"microblogger u_f is not observed"*. The set of tweets published by microblogger u_f is defined by $T(u_f)$.

As introduced in section 6.2 in the previous chapter, microbloggers social network is represented by a multigraph $G := (U, E, \Sigma_E, \ell_E, w)$ where the set of edges E represents followership, retweeting and mentioning associations between microbloggers. A social importance may be associated to each microblogger as discussed previously.

Period: Let O be the set of periods. A period $o_e \in O$ corresponds to a time window with a duration Δt. Each period covers a temporal interval defined by $[\theta_{o_e}, \theta_{o_e} + \Delta t]$ with timestamp θ_{o_e} defines the start time of period o_e and $\theta_{o_e} + \Delta t$ respective end time. A random variable $o_e \in \{0, 1\}$ is associated to each period. The event $o_e = 1$, noted o_e, denotes *"the period o_e is selected"*. Conversely, $o_e = 0$, noted $\bar{o}_e$, denotes *"the period o_e is not selected"*.

We notice that successive periods can not be parallel or overlapped $\theta_{o_{e+1}} - \theta_{o_e} >= \Delta t$. Furthermore, timestamps values are ranked in increasing order. Accordingly, period o_e with $\theta_{o_e} = 1$ is anterior to period o_{e+1} defined by $\theta_{o_{e+1}} = 2$.

7.2 The Inference Bayesian network based model for tweet search

We introduce in this section our Bayesian network model for tweet search. In particular, this model integrates, within an inference Bayesian network, several relevance features including text similarity measures, microblogger influence, presence of hashtags and temporal magnitude. These features are modeled using four types of nodes that represent queries, terms, tweets and microbloggers. With regard to conditional probabilities that involve all these types of nodes, the relevance of tweets is interpreted as a joint probability of observing both query and tweets. For this aim, we will first describe the topology of the Bayesian inference network for tweet search then we focus on the query evaluation process.

7.2.1 Network topology

The Bayesian network for tweet search is represented by a graph $G(X, E)$, where nodes $X = Q \cup K \cup T \cup U$ correspond to the set of random variables and the set of edges $E = X \times X$ represents conditional dependencies among them. Q, K, T and U correspond, respectively, to the sets of queries, terms, tweets and microbloggers nodes. The set of nodes and edges are defined in the following.

7.2.1.1 Information nodes

Inference network nodes represent random variables of the Bayesian model. In accordance with random variable types, nodes are classified in homogenous layers. Figure 7.1 shows partition of layers and interconnection between each others.

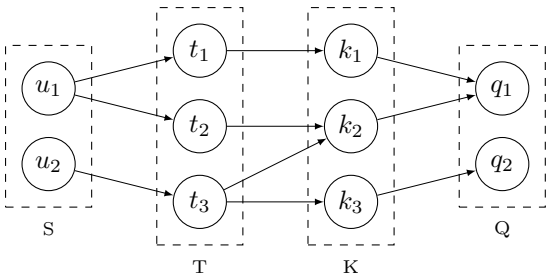


Figure 7.1: The inference Bayesian network model for tweet search

Inference Bayesian network model for tweet search consists on the four following interposed layers:

Query layer Q: A user query q is associated to a random variable $q \in Q$ and respectively a query node q in query layer. These nodes represent the root of inference the Bayesian network.

Terms layer K: A term k_i is associated to a random variable $k_i \in K$ and respectively a term node k_i in terms layer. In practice, only terms from the query are represented in terms layer. Other terms are assumed not effective for tweet relevance and subsequently ignored in the Bayesian network structure. Terms layer is interposed between query layer and tweets' layer.

Tweets layer T: A term t_i is associated to a random variable $t_j \in T$ and respectively a tweet node t_j in tweets layer. Tweets represented in this layer include at least one term from terms layer K. Other tweets are assumed irrelevant to the query. Tweets layer is interposed between terms' layer and microbloggers' layer.

Microbloggers layer U: A microblogger u_f is associated to a random variable $u_f \in U$ and respectively a microblogger node u_f in microbloggers' layer. Only microbloggers having published at least one tweet from the instantiated tweet in tweets layer are considered. Nodes of microbloggers layer represent leaf nodes of the Bayesian network model.

7.2.2 The information edges

Edges of the Bayesian network express conditional dependencies between random variables. Three types of edges are identified in the inference network model for tweet search.

Term to query: Edges connecting query $q \in Q$ with all parent terms $k_i \in K$ represent the chance of generating the query from connecting terms.

Tweet to term: A term k_i is connected to parent tweets $t_j \in T$ where it is present. Edges from tweets to terms show the chance of observing a particular term in the tweet.

Microbloggers to tweet edges A tweet node $t_i \in T$ is connected to a one parent node corresponding to the microblogger who has published t_j. This edge shows that the event of observing a tweet t_j with regards to microblogger u_k. To avoid cycles in the graph, microbloggers are presumed mutually independent.

7.2.3 Query evaluation

.

The relevance of tweet t_j considering query q submitted at θ_q is assimilated to the joint probability that both events $t_j = 1$ and $q = 1$ occur. Accordingly, tweet relevance is defined by probability $P(q \wedge t_j|\theta_q)$. In order to respect the temporal constraint in tweet search, we filter all the tweets with corresponding date θ_{t_j} is posterior to query date θ_q. We set relevance probability to $P(t_j|q) = 0$ for each tweet t_j where $\theta_{t_j} > \theta_q$. For the rest of tweets, this probability is written as:

$$P(q \wedge t_j) = \sum_{\vec{k}} P(q|\vec{k})P(\vec{k}|t_j)P(t_j) \tag{7.2}$$

with $\vec{k}$ refers to query term configurations.

Assuming term independence, the probability of observing term configuration $\vec{k}$ having tweet t_j is written as:

$$P(\vec{k}|t_j) = \prod_{\forall i|on(i,\vec{k})=1} P(k_i|t_j) \times \prod_{\forall i|on(i,\vec{k})=0} P(\bar{k_i}|t_j) \tag{7.3}$$

Furthermore, the probability of observing tweet $P(t_j)$ depends on the respective microblogger u_k. Probability $P(t_j)$ is therefore computed as:

$$P(t_j) = P(t_j|u_k) \times P(u_k) \tag{7.4}$$

Substituting $P(\vec{k}|t_j)$ and $P(t_j)$ in equation 7.2, the relevance of a tweet t_j with regards to query q is finally computed as:

$$P(q \wedge t_j) = \sum_{\vec{k}} P(q|\vec{k}) \times P(t_j|u_k) \times P(u_k) \times \left(\prod_{\forall i|on(i,\vec{k})=1} P(k_i|t_j) \times \prod_{\forall i|on(i,\vec{k})=0} P(\bar{k_i}|t_j) \right) \tag{7.5}$$

To deal with the query time, we propose to filter tweets with respect of query time θ_q. Hence, posterior tweets are discarded from the result set. The relevance probability of tweet t_j with regard to query q submitted at time θ_q is therefore computed as follows:

$$RSV(q, t_j, \theta_q) = \begin{cases} P(q \wedge t_j), & \text{if } \theta_{t_j} \leq \theta_q \\ 0, & \text{otherwise} \end{cases} \tag{7.6}$$

with θ_{t_j} corresponds to the publishing time of tweet t_j.

7.2.4 Probability estimation

We focus in what follows on the conditional probabilities introduced in equation 7.5 and we present corresponding computing formulas.

7.2.4.1 Computing probability $P(q|\vec{k})$

The probability $P(q|\vec{k})$ of observing the query q having the parent configuration $\vec{k}$ helps to weight the different combinations of the query terms. We estimate the probability of query q with m parent terms $\{k_1, k_2, \ldots, k_m\}$ as follows:

$$P(q|\vec{k}) = p_1 \times p_2 \times \cdots \times p_m \tag{7.7}$$

with $p_i = on(i, \vec{k})$.

We notice that $P(q|\vec{k}) > 0$ only if all query terms are positively instantiated in the query parent configuration $\vec{k}$. This does not discard tweets containing partial terms of the query but gives an absolute importance to the query parent configuration where all the terms are instantiated.

7.2.4.2 Computing probability $P(k_i|t_j)$

The probability $P(k_i|t_j)$ of observing term k_i in tweet t_j depends, on the one hand, on the term's occurrence and on the other hand on the tweet's properties. This probability is computed using the term frequency $F(k_i, t_j)$, the hashtag presence $H(k_i, t_j)$, the time magnitude $T(k_i, t_j)$ and the tweet length $L(t_j)$. We notice that $F(k_i, t_j)$ and $H(k_i, t_j)$ address the term's occurrence while $L(t_j)$ and $F(k_i, t_j)$ address the tweet's properties. The probability $P(k_i|t_j)$ is defined by:

$$P(k_i|t_j) = (1-\mu)F(k_i, t_j)\ H(k_i, t_j) + \mu\ T(k_i, t_j)\ L(t_j) \tag{7.8}$$

$$P(\bar{k_i}|t_j) = 1 - P(k_i|t_j) \tag{7.9}$$

with $\mu \in [0..1]$ is a smoothing parameter and the closer μ is to 0, a higher importance is given to term's appearance rather than the tweet's properties. We note that in the case where the term is not present, a default probability is assigned to the tweet depending on its length and time magnitude.

Functions introduced in equation 7.8 compute a relevance probability for each feature. These functions are detailed in what follows.

Term frequency $F(k_i, t_j)$. Due to the limited tweet length, a given term is almost used once in the same tweet. Repeating the term will emphasize it but don't attribute it an absolute highlight compared to other terms naturally occurring once. We propose so to substitute the common *tf* measure with a graduated function $F(k_i, t_j)$ that maps high frequencies into a small interval as follows:

$$F(k_i, t_j) \begin{cases} \frac{tf_{k_i,t_j} - \beta}{tf_{k_i,t_j}}, & \text{if } k_i \text{ is present in } t_j \\ 0, & \text{otherwise} \end{cases} \tag{7.10}$$

with $\beta \in [0..1]$ and tf_{k_i,t_j} is the frequency of term k_i in tweet t_j.

Hashtag score $H(k_i, t_j)$. Marking term k_i with a hashtag $\#k_i$ would put a highlight on it and increases its importance compared to the other terms in the tweet. We propose a hashtag function $H(k_i, t_j)$ that leverages the importance of the terms as follows:

$$H(k_i, t_j) \begin{cases} 1 - \frac{h}{tf_{\#k_i,t_j}}, & \text{if } \#k_i \text{ is present in } t_j \\ h, & \text{otherwise} \end{cases} \tag{7.11}$$

with $h \in [0..0.5]$ is the default hashtag score and $tf_{\#k_i,t_j}$ is the frequency of the hashtag $\#k_i$ in tweet t_j. We note in equation 7.11 that $H(k_i, d_j) \geq (1-b)$ if the hashtag $\#k_i$ occurs a least once in tweet t_j.

Time magnitude $T(k_i, t_j)$. The term's importance varies over the time, increasing and decreasing with its presence in the tweet feeds. Therefore, the probability of observing term k_i depends also on the time when tweet t_j is submitted. This probability would be more important when term k_i is frequently used. Considering term k_i, we measure the time magnitude of tweet t_j as follows:

$$T(k_i, t_j) = 0.5 + 0.5\frac{df_{k_i,\Gamma_j}}{|\Gamma_j|} \tag{7.12}$$

$$\Gamma_j = \left\{t_k, |\theta_{t_j} - \theta_{t_k}| \leq \Delta t\right\} \tag{7.13}$$

with Γ_j refers to the set of temporal neighbors of tweet t_j within the $2\Delta t$ time window. df_{k_i,Γ_j} is the number of tweets in Γ_j containing term k_i.

Tweet length $L(t_j)$. Very short tweets are considered ambiguous and pointless. Unlike common measures maximizing the scores of short documents, we propose to favour tweets closer to the average tweets length avg_{tl}. A length score $L(t_j)$ is assigned to each tweet as follows:

$$L(t_j) = \frac{1}{1 + \left|avg_{tl} - tl_{t_j}\right|} \tag{7.14}$$

7.2.4.3 Computing probability $P(t_j|u_k)$.

The probability $P(t_j|u_k)$ of arriving to tweet t_j having a microblogger u_k weights the different tweets of one microblogger. Considering the set $\mathcal{T}_{u_k}$ of instantiated tweets published by the microblogger u_k, this probability is computed as follows:

$$P(t_j|u_k) = \frac{1}{|\mathcal{T}_{u_k}|} \tag{7.15}$$

7.2.4.4 Computing probability $P(u_k)$.

Since microbloggers are root nodes, a prior probability $P(u_e)$ are assigned with respect to their social importance. A microblogger would receive a high importance if he plays a key role in his social network. In this work, the social importance is linked to influence, leadership and conversional activity as discussed in chapter 6. Accordingly, our proposed algorithms *InfRank*, *LeadRank* and *DicussRank* could be used in this context to evaluate the social importance of microbloggers.

We reiterate that the social network of microbloggers is modeled by a multigraph $G := (U, E, \Sigma_E, \ell_E, w)$ where U represents the set of microbloggers and $R = U \times U$ denotes the set edges representing followerships, retweets and mentioning relationships. However, to avoid the dominance of some celebrities characterized by a high retweet number, we propose to apply social ranking algorithms only on

the sub-network of microbloggers generated by the instantiated retweets. This helps to evaluate microbloggers' influence regarding to a specific topic.

In practice, the probability $P(u_f)$ of observing microblogger u_f is defined for instance by leadership score $Ldr^k(u_f)$ computed by *LeadRank* algorithm.

$$P(u_f) = Ldr^k(u_f) \tag{7.16}$$

Other social ranking algorithms presented in chapter 6 could be used instead.

7.3 The belief Bayesian network based model for tweet search

The inference Bayesian network model for tweet search provides a unified framework that models several relevance features. However, this model presents a limited topology. The integration of other nodes such as periods is not allowed within the interposed structure of network layers. Inspired by work of de Cristo et al. (2003) that proposes to integrate topical and hyperlink-based authority evidences into a Bayesian belief network, we propose here a second tweet search model that integrated topical, social and temporal evidences for tweet relevance within a belief Bayesian network. In particular, the relevance of a tweet with regard to a query is estimated with respect of term occurrence, the social importance of a microblogger and term distribution over time. We introduce in what follows the topology of the belief Bayesian network for tweet search and then we focus on the query evaluation process.

7.3.1 Network topology

Figure 7.2 presents the topology of our Bayesian network model for tweet search. Unlike previous inference Bayesian network model where the queries represent network roots, terms are positioned here as the root of the belief network. This allows to model query and tweets are modeled in two separated layers and integrate additional sources of evidence in each layer.

In addition to random variables t_j that model the probability of observing the tweet in inference Bayesian networks, a tweet is represented here by three other random variables t_{kj}, t_{sj} and t_{oj}. First variable t_{kj} models the event of observing tweet t_j given an implicit knowledge of term occurrence in the tweet. The random variable t_{sj} models the event of observing t_j given an implicit knowledge of microblogger social influence. Finally, the random variable t_{oj} models the event of observing tweet t_j given an implicit knowledge of the time magnitude of tweet. These probabilities decompose the event of observing the tweet into three evidences: topical evidence, social evidence and temporal evidence.

The Bayesian network model for tweet search contains three connected networks:

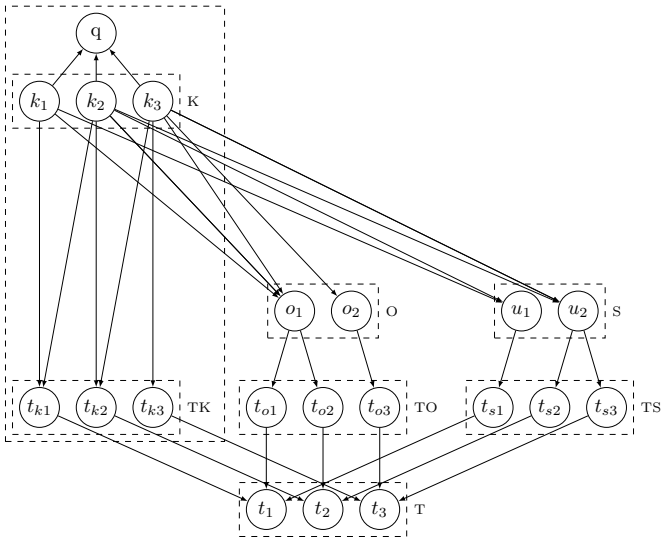


Figure 7.2: Belief network model for tweet search

Tweet network: Each term k_i in the Bayesian network is represented with a node. The set of term nodes constitutes the term layer K. A user query is modeled with node q. It exists a directed edge (k_i, q) from the query node to parent term k_i if only $on(k_i, q) = 1$. A tweet t_j is represented at first time by three nodes t_{kj}, t_{sj} and t_{oj}. Respectively, these nodes belong to the topical evidence layer TK, the social evidence layer SO and the temporal evidence layer TS. t_{kj} is the only tweet node connected directly to term nodes. An edge (k_i, t_{kj}) connects tweet t_{kj} to each included term k_i if $on(k_i, t_j) = 1$. t_{kj}, t_{sj} and t_{oj} are connected to a another node t_j. For $x \in \{k, s, o\}$, it exists an edge (t_{xj}, t_j) from t_{xj} to t_j. The set of nodes t_j constitutes the tweet layer T.

Microblogger network: Each microblogger u_f is represented by a node. These nodes constitute the social layer S. Microbloggers nodes are connected to correspondent tweet nodes in the social evidence layer TS. An edge (u_f, t_{sj}) is defined between a microblogger u_f and a tweet node t_{sj} if the tweet t_j is published by u_f. We notice that tweet t_v and a respective retweet t_w are represented by two independent nodes. In this case, retweet node t_w is connected to the retweeting microblogger instead of the original author of tweet t_v. In addition, microbloggers are connected to term nodes in layer K. An edge (k_i, o_e) connects a microblogger u_f to each term k_i appeared in one of his tweet at least $\{k_i \in K, (u_f, t_{sj}) \in E \wedge on(k_i, t_j) = 1\}$.

Period network: Each period o_e is represented by a node. Period nodes constitute the temporal layer O. Periods are connected to nodes from tweet temporal

layer TO and term layer K. An edge (o_e, t_{oj}) connects a period o_e to a tweet node t_{oj} if t_j is published in the respective time window $\theta_{o_e} - \theta_{t_j} \leq \Delta t$. Since periods are not overlapped, a tweet is connected to one only period. Besides, a node o_e is connected to each term node k_i observed in the respective period $\{k_i \in K, on(k_i, t_j) = 1 \wedge \theta_{o_e} - \theta_{t_j} \leq \Delta t\}$.

7.3.2 Query evaluation

The relevance of tweet t_j with respect to query q submitted at time θ_q is computed by the probability $P(t_j|q, \theta_q)$. Ignoring the query date, this probability is estimated by:

$$P(t_j|q) = \frac{P(t_j \wedge q)}{P(q)} \tag{7.17}$$

$P(q)$ have a constant value for all the tweets. $P(t_j|q)$ is then approximated with $P(t_j|q) \propto P(t_j \wedge q)$. Based on the topology of the Bayesian network for tweet search, the probability $P(t_j|q)$ is developed as follows:

$$P(t_j|q) \propto \sum_{\vec{k}} P(q|\vec{k})P(t_j|\vec{k})P(\vec{k}) \tag{7.18}$$

with $\vec{k}$ is a term configuration.

To simplify the computation of probability $P(t_j|q)$, only instantiated terms in the query are considered in the configuration $\vec{k}$.

In fact, the probability $P(t_j|\vec{k})$ depends on thee sources of evidence: topical evidence, social evidence and temporal evidence. This probability $P(t_j|\vec{k})$ is rewritten as follows:

$$P(t_j|\vec{k}) = P(t_{kj}|\vec{k})P(t_{sj}|\vec{k})P(t_{oj}|\vec{k}) \tag{7.19}$$

By substituting $P(t_j|\vec{k})$ in formula 7.18, tweet relevance is estimated as:

$$P(t_j|q) \propto \sum_{\vec{k}} P(q|\vec{k})P(t_{kj}|\vec{k})P(t_{sj}|\vec{k})P(t_{oj}|\vec{k})P(\vec{k}) \tag{7.20}$$

7.3.3 Computing conditional probabilities

We detail in what follows the computation of the conditional probabilities in equation 7.20.

7.3.3.1 Probability $P(\vec{k})$

The probability $P(\vec{k})$ corresponds to the likelihood of observing term configuration $\vec{k}$. We assume that all the configurations are independent and have an uniform probability to be observed. Let n be the query length which corresponds also to the number of terms represented in configuration $\vec{k}$, the probability $P(\vec{k})$ is estimated as:

$$P(\vec{k}) = \frac{1}{2^n} \tag{7.21}$$

7.3.3.2 Probability $P(q|\vec{k})$

The probability $P(q|\vec{k})$ determines the likelihood of generating query q from a term the configuration $\vec{k}$. For this aim, we first propose to weight each term k_i according to its appearance in the collection $w_{k_i} = \frac{df_{k_i}}{N}$ with df_{k_i} is the number of tweets containing term k_i and N is the number of posterior tweets to query q. According to the set of positively instantiated terms in configuration $c(\vec{k})$, the probability $P(q|\vec{k})$ is computed using the Noisy-Or operator:

$$P(q|\vec{k}) = \begin{cases} \frac{1 - \prod_{k_i \in c(\vec{k}) \wedge q} w_{k_i}}{1 - \prod_{k_i \in q} w_{k_i}}, & \text{if } c(\vec{k}) \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \tag{7.22}$$

Thus, configurations with significant terms in the collection are highlighted in contrast of configurations that present commonly used terms or stopwrods.

7.3.3.3 Probability $P(t_j|\vec{k})$

The probability $P(t_j|\vec{k})$ that tweet t_j is generated by configuration $\vec{k}$ measures the topical similarity between the tweet and the configuration. This probability could be estimated based on the term frequency tf_{k_i,t_j}. However, terms have less chance to be repeated once tweet length is limited. Similarly to the inference Bayesian network model, we propose to weight each term k_i as follows:

$$w_{k_i,t_j} = \begin{cases} \frac{tf_{k_i,t_j} - \beta}{tf_{k_i,t_j}}, & \text{if } on(k_i, t_j) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{7.23}$$

where tf_{k_i,t_j} is the frequency of term k_i in tweet t_j.

w_{k_i,t_j} maps high frequencies into a small interval. We note that small values of β reduces the weight of frequent terms. Accordingly, we give less importance to term frequency rather than term presence in the case of long queries.

The probability $P(t_j|\vec{k})$ is finally computed as:

$$P(t_j|\vec{k}) = \begin{cases} \sigma \sum\limits_{k_i \in c(t_j) \wedge c(\vec{k})} w_{k_i,t_j}, & \text{if } c(t_j) \wedge c(\vec{k}) \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \tag{7.24}$$

where σ is a normalization factor defined by:

$$\sigma = \frac{1}{\sum\limits_{\vec{k}} \sum\limits_{k_i \in c(t_j) \wedge c(\vec{k})} w_{k_i,t_j}} \tag{7.25}$$

7.3.3.4 Probability $P(t_{sj}|\vec{k})$

The probability $P(t_{sj}|\vec{k})$ of observing tweet t_j having the social influence of corresponding microbloggers and term configuration $\vec{k}$ is estimated as follows:

$$P(t_{sj}|\vec{k}) = P(t_{sj}|u_f)P(u_f|\vec{k}) + P(t_{sj}|\bar{u}_f)P(\bar{u}_f|\vec{k}) \tag{7.26}$$

The probability $P(t_{sj}|\bar{u}_f)$ of observing the tweet while corresponding microblogger u_f is not observed, is equal to 0. The probability $P(t_{sj}|\vec{k})$ is therefore transformed to:

$$P(t_{sj}|\vec{k}) = P(t_{sj}|u_f)P(u_f|\vec{k}) \tag{7.27}$$

Assuming that the two events of observing microblogger u_f and configuration $\vec{k}$ are independent, we write:

$$P(t_{sj}|\vec{k}) = P(t_{sj}|u_f)P(u_f) \tag{7.28}$$

First, the probability $P(t_{sj}|u_f)$ of observing tweet t_j having the microblogger u_f weights the tweets of each microblogger. This probability is computed equally for set of tweets $T(u_f)$ published by microblogger u_f.

$$P(t_{sj}|u_f) = \frac{1}{|T(u_f)|} \tag{7.29}$$

Similarly to the inference Bayesian network model, the prior probability $P(u_f)$ of observing microblogger u_f is defined by the social importance of the social network. For this aim, we propose to extract the social network of microbloggers from instantiated microbloggers with respect of the social network structure defined in chapter 6. In particular, the social network is defined by a mutligraph where microbloggers are connected with followership, retweet, and mentioning social relationships.

The social importance of microbloggers is determined according to his influence, leadership or discussion activities. Previously defined algorithms *InfRank*,

LeadRank and *DisuccsRank* may be used in this context to estimate the importance of microbloggers in the social network. As proposed before, the $P(u_f)$ is computed for example as the leadership score of microbloggers in the social network.

$$P(u_f) = Ldr^k(u_f) \tag{7.30}$$

7.3.3.5 Probability $P(t_{oj}|\vec{k})$

The probability $P(t_{oj}|\vec{k})$ of observing tweet t_j knowing period o_e and term configuration $\vec{k}$ is estimated as follows:

$$P(t_{oj}|\vec{k}) = P(t_{oj}|o_e)P(o_e|\vec{k}) + P(t_{oj}|\bar{o}_e)P(\bar{o}_e|\vec{k}) \tag{7.31}$$

The probability of observing the tweet outside the respective period $P(t_{oj}|\bar{o}_e)$ is equal to 0. Thus, $P(t_{oj}|\vec{k})$ is written as:

$$P(t_{oj}|\vec{k}) = P(t_{oj}|o_e)P(o_e|\vec{k}) \tag{7.32}$$

The probability $P(t_{oj}|o_e)$ of observing tweet t_j, having the period o_e, weights the different tweets published in this period. The visibility of a tweet increases with the number of received retweets. Consequently, this probability is computed proportionally to the number of retweets generated by tweet t_j in the same period.

$$P(t_{oj}|o_e) = \frac{1 + |R(t_j, o_e)|}{|T(o_e)|} \tag{7.33}$$

where $R(t_j, o_e)$ is the set of retweets of tweet t_j in period o_e. $T(o_e)$ is the set of tweets published in period o_e.

The probability $P(o_e|\vec{k})$ of selecting period o_e, having configuration $\vec{k}$, weights the different periods. We estimate this probability based on two factors. First, we consider the time decay between period o_e and query date θ_q. In fact, recent tweets are more likely to interest microblog users. Second, we consider the percentage of tweets published in o_e and containing the configuration $\vec{k}$. This highlights active period of the configuration $\vec{k}$ that concurs with a real world event. Periods are weighted as followings:

$$P(o_e|\vec{k}) = \frac{\log(\theta_q - \theta_{o_e})}{\log(\theta_q - \theta_{o_s})} \times \frac{df_{\vec{k},o_e}}{df_{\vec{k}}} \tag{7.34}$$

with θ_q, θ_{o_e} and θ_{o_s} are respectively the timestamps of query q, the period o_e and the period o_s when the oldest tweet containing the term configuration $\vec{k}$ is published with $\theta_{o_s} \leq \theta_{o_e} \leq \theta_q$. $df_{\vec{k},o_e}$ is the number of tweets published in o_e and containing configuration $\vec{k}$. $df_{\vec{k}}$ is the number of tweets with a term configuration $\vec{k}$.

7.4 Experimental evaluation

We conduct a series of experiments on TREC Microblog dataset *tweet2011* in order to study the effectiveness of our Bayesian network models for tweet search. We focus in this study on the query level and we analyze the impact of each integrated feature on the retrieval performances. In these experiments, we refer to our inference Bayesian network model for tweet seach as *BNTSi* (Damak et al., 2011) and to the belief network model as *BNTSb* (Ben Jabeur et al., 2012a).

7.4.1 Experimental setup

Tweet and query dataset. These experiments are carried out using TREC 2011 and 2012 Microblog Track (Ounis et al., 2011; Soboroff et al., 2012). The tweet dataset includes about 16 million tweets published over 16 days. Table 7.1 presents general statistics about the collection. We observe that 0.07% of tweets in the collection are retweets. Mentioning tweet represent 0.45% of total tweets. We notice that this dataset is built based on Twitter API which provides a representative sample of 1% of the tweet stream (McCreadie et al., 2012). Other tweets published in the same period are not included in the collection.

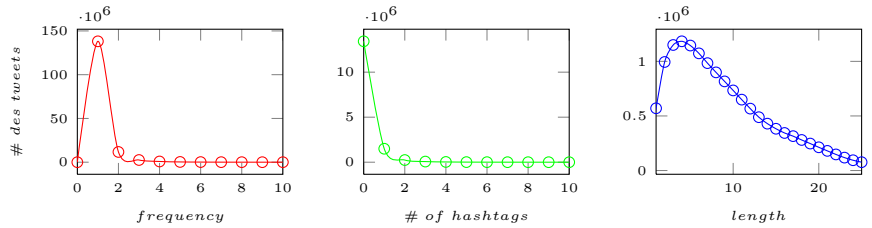
Tweet and query dataset. These experiments are carried out using TREC 2011 and 2012 Microblog Track (Ounis et al., 2011). The tweet dataset includes about 16 million tweets published over 16 days. Table 7.1 presents general statistics about the collection. We observe that 0.07% of tweets in the collection are retweets. Mentioning tweet represent 0.45% of total tweets. We notice that this dataset is built based on Twitter API which provides a representative sample of 1% of the tweet stream (McCreadie et al., 2012). Other tweets published in the same period are not included in the collection.

Tweets	16141812	Microbloggers	5356432
Retweets	1128179	Network nodes	5495081
Mentions	7193656	Network retweets	1061989
Terms	7781775	Network mentions	9503013

Table 7.1: Dataset statistics

We extracted the social network from the tweets in the dataset. About 5.3 million microbloggers are found. We notice that the number of network nodes exceeds the number of microbloggers inside the collection as presented in table 7.1. This is explained by the fact that some retweets and mentions point to other users outside the collection. Each microblogger in the network is involved in about 0.19 retweet associations and 1.73 mention associations.

Figure 7.3 presents the distribution of term frequency, hashtags and document length. Figure 7.3.(a) shows that terms are often used once in the same tweet. Figure 7.3.(b) shows that only 2% of tweets contains 2 or more hashtags while the majority of tweets (88%) don't include any hashtag. Figure 7.3.(c) shows that the length distribution presents a peak at 4. We also report that 53% of tweets include 8 terms at least with an average tweet length estimated to 11.05.



(a) Term frequency (b) Hashtags (c) Tweet length

Figure 7.3: Term frequency, hashtags and length distributions

Real-time ad-hoc task. The real-time ad-hoc task of TREC Microblog includes 49 topic for 2011 edition and 60 topic for 2012. Each topic contains a query and a respective timestamp when the query is submitted. $P@30$ is reported as the official measure in both editions. In 2011 and in contrast of other TREC tasks where results are ranked by score, real-time search task ranks results by the inverse chronological order. This constraint was dropped in 2012 where tweets are finally ranked according to their relevance score. In addition, Mean Average Precision *MAP* and *ROC* curves were referenced as non official measures for TREC Microblog 2011 and 2012, respectively. In order to respect temporal constraints, we do not integrate any future data or external resource. All term counts are computed using anterior tweets to the query time including tf_{k_i} that counts the number of tweets that containing term t_j.

Tweet filtering. In order to enhance the precision of our model, we apply the next filter to the final result set.

- *Language filter:* Tweets in other languages except English are ignored in the assessment process. In view of that, we keep only English tweet in the result set. To detect tweet language we used a Naive Bayes classier[2] that recognizes about 50 written languages.
- *Retweet filter:* Retweets are presumed irrelevant in this track even though they discuss the query topic. Accordingly, we remove from the result set all

[2]http://code.google.com/p/language-detection/

the tweets starting with *RT @username* including native retweets as well as non-conventional retweets.

– *Reply filter:* We propose also to remove all the replies starting with *@username* as discussion tweets would be irrelevant for this task.

7.4.2 Model tuning

In order to tune β and Δt parameters of our model, we conduct in what following empirical experiments using 2011 query dataset of TREC microblog Track. In particular, we study the impact of β and Δt parameters on the $P@30$.

Figure 7.4 presents $P@30$ precisions for different values of β obtained by considering only term frequency in our two Bayesian network models respectively in equations 7.23 and 7.10. All the other features are deactivated for this purpose.

For $\beta = 0$, term frequency tf_{k_i,t_j} is assimilated to Boolean frequency that simply indicates the presence of a term in the tweet. The closer β value is to 1, the largest is the interval of frequency mapping and thus high frequencies are emphasized. $P@30$ attains a maximal value with $\beta = 0.5$. Beyond this value, a significant decrease of model performances is observed to attain a negative change of -60% for $\beta = 1$. These results as well term frequency distribution presented in figure 7.3 confirm that the high frequency of a term is less significant than the simple occurrence in the tweet. However, reducing the gap between high frequencies and Boolean occurrence may improve the retrieval performances.

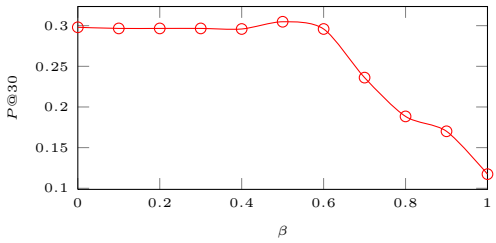


Figure 7.4: Tuning parameter β

Figure 7.5 shows the impact of time window Δt on the search effectiveness. For this aim, $P@30$ is compared for different Δt values of our Bayesian inference model *BNTSb* where only the topical and the temporal features are activated. Precisions $P@30$ achieve a maximal performances at $\Delta t = 1\ day$. Outside the interval $[4h, 30h]$, a significant precision decrease is observed for configuration *BTNSi.KO*. This interval correlates with major updates by news channels and newspapers.

For the next experiments, we set β to 0.5 and $\Delta t = 1day$.

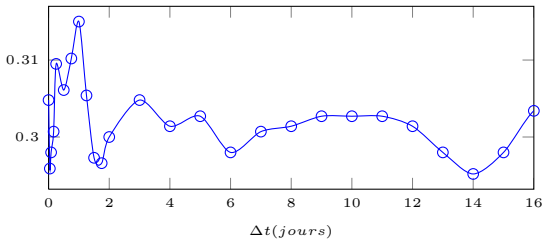


Figure 7.5: Tuning parameter Δ

7.4.3 Comparing social ranking algorithms

The probability $P(u_f)$ of observing microblogger u_f expresses the social importance of the microblogger. This probability may be computed using one of our microblogger ranking algorithms presented in chapter 6, namely *InRank*, *LeadRank* and *DiscussRank*. We compare in what follows the impact of these ranking algorithms on the effectiveness of our two Bayesian network models for tweet search *BNTSi* and *BNTSb*.

Figure 7.6 presented MAP values obtained by 3 social ranking algorithms. We notice that only the topical and the social components of our two Bayesian network models *BNTSi* and *BNTSb* are activated. *MAP* values are compared for both 2011 and 2012 query datasets of TREC Microblog Track.

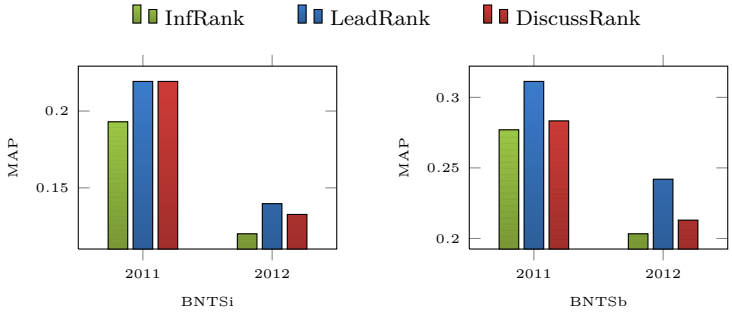


Figure 7.6: Comparing social raking algorithms

Results shows that *InRank* presents lower *MAP* values. This is explained by the fact that retweets are presumed irrelevant and thus ignored in relevance assessment process. Furthermore, there's a few chance that the original tweet is included in the dataset as only 1% of tweet stream is tracked. In view of that, a microblogger must have published two tweets in the query topics at least to be identified by *InRank* as network influencers.

LeadRank and *DiscussRank* show however better results rather than $InfRank$. *LeadRank* presents almost best MAP precision for both topic sets 2011 and 2012

to attain an improvement of 19% compared to *InRank* and improvement of 14% compared to *DiscussRank* in the case of *BTNSb* and TREC 2012 query dataset. In fact, *LeadRank* and *DiscussRank* show better results because they integrate mentioning relationships in addition to retweets. These relationships are more common in the social network. We thus conclude that the density of the social network has an impact on the effectiveness of the social ranking algorithms. Although they present satisfying results for microblogger ranking as shown in chapter 6, the real effect of the ranking algorithms on tweet search is not clearly accessible here due the TREC microblog evaluation specificity.

7.4.4 Evaluating retrieval effectiveness

We compare our Bayesian Network models for Tweet Search based on Bayesian inference network *BNTSi* and Bayesian belief network *BNTSb* to some representative models from TREC Microblog Track. A brief description of these models is presented in table 7.2.

Notation	Year	Rank	Description
isiFDL	2011	1^{st}	Learn to Rank model based on *Markov Random Field* model (Metzler and Cai, 2011)
DFReeKLIM	2011	2^{nd}	*Kullback-Leibler* based model (Amati et al., 2011)
KAUSTRerank	2011	17^{th}	Learn to Rank model that considers user authority (Jiang et al., 2011). Basic run is noted *KAUST-Base*
gust	2011	20^{th}	Language model that considers the query temporal profile (Efron, 2011)
Disjunctive	2011		Official track baseline that ranks tweets containing at least one term on the query by inverse chronological order
hitURLrun3	2012	1^{st}	Query and document expansion model that takes into account included URL (Han et al., 2012)
uwatgclrman	2012	2^{nd}	Manual trained feedback with a logistic regression classier (Roegiest and Cormack, 2012)
hitLRrun1	2012	3^{th}	Learning to rank that considers text-based features, non-text features (e.g., URL, hashtag, etc) and user features (i.e., followers count) (Han et al., 2012)
ICTWDSERUN1	2012	4^{th}	Pseudo relevance feedback based on *indri* retrieval system (Zhu et al., 2012)

Table 7.2: Representative models from TREC Microblog track

Table 7.3 presents a comparison of $P@30$ and MAP with different thresholds on the result set size (*cutoff*). First, we note that the threshold choice impacts the retrieval effectiveness. In fact, time-ranked result set presents low error risks if only some few tweets are included and vice versa. One optimal choice to improve $P@30$ precision is to return only the top 30 tweets for each query. That

is the case of the first and the second official runs *isiFDL* and *DFReeKLIM*. However such evaluation oriented method presents lower results for long result sets typically for *MAP* precision at threshold 1000 as shown by *DFReeKLIM*. Conversely, presenting longer results sets may enhance the performance of some systems. For instance, *gust* run presents slightly better *P*@30 results at threshold 300 compared to threshold 30. As this system investigates the temporal distribution of top selected tweets, the accuracy of the temporal query profile may be better for longer result sets.

		Cutoff	*P*@30		*MAP*	
isiFDL	*	30	**0.4551**	(−25%)	0.2439	(−27%)
DFReeKLIM	*	30	0.4401	(−22%)	**0.2811**	(−37%)
BNTSb		30	0.3422		0.1774	
BNTSi		30	0.3047	(+12%)	0.1542	(+15%)
gust	*	30	0.3218	(+6%)	0.1812	(−2%)
Median	*		0.2575	(+33%)	0.1426	(+24%)
KAUSTRerank	*	50	**0.3456**	(−9%)	**0.2390**	(−17%)
KAUSTBase	*	50	0.3347	(−7%)	0.1902	(+5%)
BNTSb		50	0.3129		0.1990	
gust		300	**0.3220**	(−31%)	0.1970	(+12%)
BNTSb		300	0.2231		**0.2201**	
BNTSb		1000	**0.1844**		**0.1929**	
DFReeKLIM	*	1000	0.1136	(+62%)	0.1651	(+17%)
Disjunctive	*	1000	0.0986	(+87%)	0.1411	(+37%)

Table 7.3: Comparison of *P*@30 and *MAP* for TREC Microblog 2011 (* official results)

This issue has revealed a critical evaluation problem of TREC Microblog Adhoc search 2011. Studying *P*@30 of time re-ranked tweets may compare the accuracy of systems but does not allow to access to the real performances of the system as its effectiveness is strongly dependent on the size of the results set. In the 2012 edition, TREC microblog organizers proposed to rank results by scores as commonly used. We will discuss 2012 results later.

As shown in table 7.3, *BNTSb* model shows better results than *BNTSi*. This is explained due the integration method that focuses on each feature namely the topical, the temporal and the social relevance factors. *BNTSb* presents an improvement of 33% compared to TREC *P*@30 median and an improvement of 24% compared to *MAP* median. A difference of about −25% is noted compared to 1^{st} model *isiFDL*. Considering the social-based models *KAUSTBase* and *KAUSTRerank*, *BNTSb* shows inferior results, expect for *KAUSTBase MAP*. We notice that this model integrates URL-based feature. Compared to time-based model *gust*, *BNTSb* presents higher *P*@30 values with the threshold set to 30. The *gust* model shows however higher *P*@30 than *BNTSb* a cutoff at 300, and vice versa for *MAP* values. Considering *P*@30 for full result set (1000 tweets), *BNTSb* presents an improvement of 37% compared to the *Disjunctive*

baseline. We note also an improvement of 17% compared to the 2^{nd} ranked model *DFReeKLIM*.

Table 7.4 compares *P*@30 and *MAP* values obtained by our two models *BNTSb* and *BNTSi* for TREC Microblog 2012. First, we note that both of our models overpasses the 4^{th} system in TREC official ranking. Once again, *BNTSb* shows higher results than *BNTSi* model. In fact, *BNTSb* outperforms first model *hitURLrun3* with an improvement of 22% for *P*@30. However, *hitURLrun3* shows higher *MAP* values.

		Cutoff	*p*@30		*MAP*	
BNTSb		30	**0.3332**		0.2466	
hitURLrun3	*	30	0.2701	(+23%)	**0.2642**	(−7%)
uwatgclrman	*	30	0.2559	(+30%)	0.2277	(+8%)
hitLRrun1	*	30	0.2446	(+36%)	0.2411	(+2%)
BNTSi		30	0.2410	(+38%)	0.1472	(+68%)
ICTWDSERUN1	*	30	0.2384	(+40%)	0.2093	(+18%)
Median	*		0.1807	(+84%)	0.1486	(+66%)

Table 7.4: Comparison of *p*@30 and *MAP* for TREC Microblog 2012 (* official result)

These results confirm that our proposed models which are based on social and temporal features are competitive on the first hand to Learning to Rank models (*hitLRrun1*) and on the second hand to pseudo relevance feedback models *hitURLrun3*, *ICTWDSERUN1* including manual runs (*uwatgclrman*).

Analyzing *BNTSib* and *BNTSb* difference from *P*@30 median per topic for TREC microblog 2011 and 2012 in figures 7.7 and 7.8, we note that the two models present similar changes for 2012 query dataset with a Person's correlation value of 0.5866. Student's t-test confirms however that results abtained by both models are significantly different from each other, respectivly, 0.0442 for 2011 and and 0.0005 for 2011

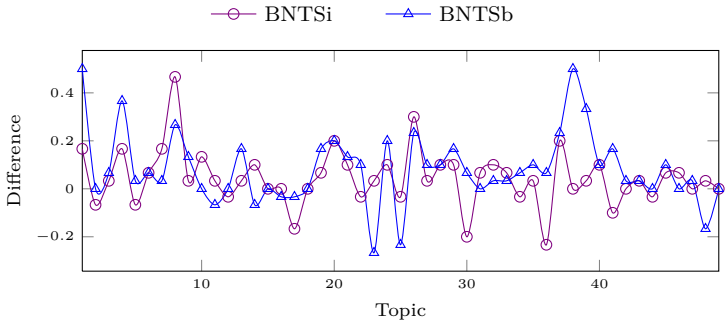


Figure 7.7: Difference to *P*@30 median for TREC Microbolog 2011

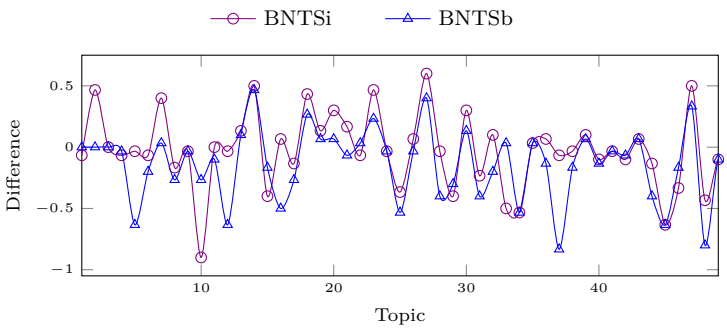


Figure 7.8: Difference to $P@30$ median for TREC Microbolog 2012

For TREC Microblog 2011, *BNTSi* realizes an improvement with regard to 32 topics out of 49. An improvement is noted for *BNTSb* with respect of 37. Conversely, *BNTSi* shows higher improvements in terms of the number of queries for TREC Microblog 2013 with 23 positive change with respect TREC median compared to only 20 improved queries for *BNTSi* . Negative difference concerns for instance topic 18 (2011) which includes only one relevant tweet. Positive difference is noted for instance for topic 1 (2011) *"BBC World Service staff cuts"* which is characterized by a high number of candidate relevant tweets with over 82500 results.

7.4.5 Feature based analysis

In order to evaluate the effectiveness of each feature with regard to the query profile we conduct at this level a series of experiments using a profiled query dataset named *"Arab Spring"*. In contrast of TREC Microblog Track, this query dataset enable to evaluate, on the first hand, the effectiveness of tweet search model in accordance to the query profile, and on the other hand, evaluate the usefulness of used feature in respect of topical, social and temporal queries. We notice that such evaluation process is not supported with TREC Microblog Track since test topics are not provided with a description field. The profile of the query could not be determined in this case. Moreover, the fact that retweets are presumed irrelevant in TREC Microblog Track would underrate the effectiveness of social information retrieval system. However, retweet is one of the main practices that reflect the social relevance of microblogs.

To build the *"Arab Spring"* dataset, we asked 2 regular Twitter users to collect 25 topics without a prior knowledge of TREC *tweets2011* corpus content. Queries are about massive democratic protests in North Africa where social media services have played a key role. Topics are classified into three categories in accordance to search motivations as claimed by Teevan et al. (2011) including social, topical and timely information needs.

Socially motivated queries aim at tracking a person's activity (*e.g.*, emph"Wael Ghonim"), find people with similar interests (*e.g.*, *"Tunisia Sidibouzid"*) and collect public sentiments about a topic (*e.g.*, *"Mubarak dissolves government"*). Topical queries aim at finding information about a specific topic (*e.g.*, *"Number of protesters in Tharir"*) or a general interest (*e.g.*, *"Tunisian revolution"*). Temporal queries aim at retrieving tweets about news (*e.g.*, *"ElBaradei arrives in Egypt"*), current events (*e.g.*, *"Clashes in Tahrir"*) and a service's status in real-time (*e.g.*, *"SMS down Egypt"*). Finally, 25 time-stamped queries are collected. Top 20 tweets returned by compared models and configurations are evaluated by 4 volunteers who are also familiarized with Twitter and interested in tracked events. The evaluation process consists on a binary judgment of relevance based on the query type and the submission time. Among 3750 analyzed tweets, 849 relevant tweets are identified for the 25 queries.

Figure 7.9 presents the $P@10$ and $P@20$ results obtained by different retrieval configurations of our *BNTSi* using *Arabic Spring'* topic dataset. A brief description of each compared configuration is presented in table 7.5.

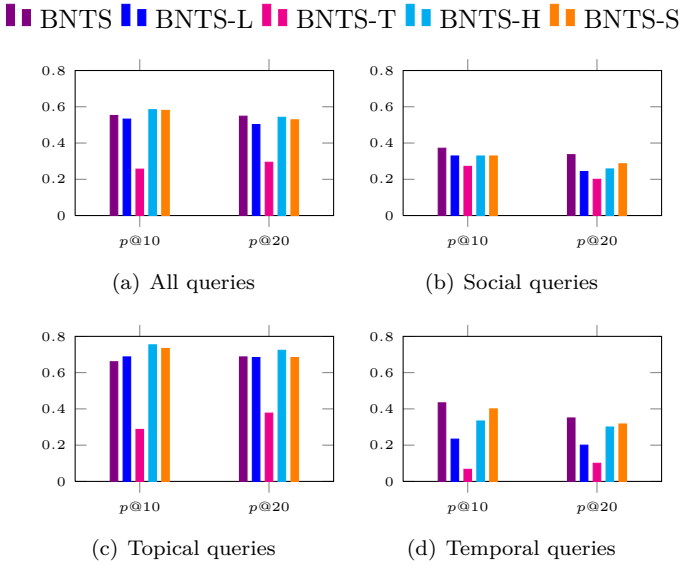
Notation	Description
BNTSi	Bayesian inference network model for tweet search
BNTSi-L	BNTSi model with *Tweet Length* feature ignored $L(t_j) = 1$
BNTSi-T	BNTSi model with *Time magnitude* feature ignored $T(k_i, t_j) = 1$
BNTSi-H	BNTSi model with *Hashtag feature* ignored $H(k_i, t_j) = b$
BNTSi-S	BNTSi model with *Social influence* feature ignored $P(u_f) = 1$

Table 7.5: Configuration discrpition of BNTSi model

Figure 7.9.(a) shows that the different configurations have close precisions except the *BNTSi-T* model presenting a considerable decline (-54%) compared to the *BNTSi* model. We conclude so that the time magnitude is a primordial feature for tweet search. The impact of the other features varies depending on the query type.

Analyzing the performance of the *BNTSi-S* model, we note a general precision decline in figures 7.9.(b) and 7.9.(d) while in figure 7.9.(c) precisions of the *BNTSi-S* model are similar or overpass their analogues for the *BNTSi* model. We conclude that the social importance of microbloggers is an important feature for social and temporal queries where users are interested in persons and fresh information. This feature is less important for topical queries where users search for specific information independently of the person who reports it.

A precision decrease is also noted for the *BNTSi-H* model in the case of social and temporal queries in contrast of topical queries where precision rises. We conclude that the hashtag feature is not helpful for specific topic search particularly if one of the query terms is frequently used as a hashtag. Considering the document length feature represented by the *BNTSi-L* model, a significant



(a) All queries (b) Social queries

(c) Topical queries (d) Temporal queries

Figure 7.9: Features impact on the tweet search

precision decrease is observed in figure 7.9.(d). This explains that users are interested in short tweets in the case of temporal queries which mainly address news and real-time information.

We present in table 7.6 a summary of the integrated features and their usefulness for tweet search. Comparison is presented by each query profile. The symbol + denotes that the respective feature is useful. Conversely, symbol − denotes that the respective feature is not useful.

	Query profile			
Feature	all	topical	social	temporal
Tweet Length	+	-	+	+
Hashtags presence	-	-	+	+
Social importance	-	-	+	+
Time magnitude	+	+	+	+

Table 7.6: Features usefulness for tweet search

Focusing on the effectiveness of social features at the query level, we analyze improvement of our *BNTSb* model with both topical and social components activated, noted *BNTSb.KS*, to the topical component of our Bayesian belief network model *BNTSb.K*. Figure 7.10 presents *MAP* difference of *BNTS.KS*

model from *BNTS.K* model.

	BNTS.K	BNTS.KS
MAP	0.1579	**0.1642**

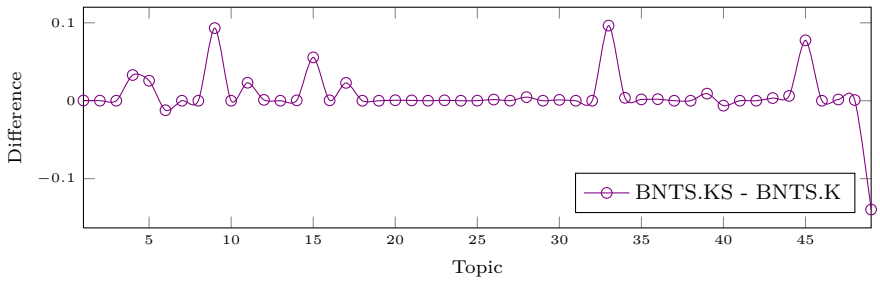


Figure 7.10: BNTS.KO difference from BNTS.K *MAP* per topic

BNTS.KS presents better results for the overall queries compared *BNTS.K*. This shows the interest of integrating the social context for tweet ranking. In particular, an important positive change is noted for instance for topic 9 *"Toyota Recall"*. In this case, relevant tweets are produced by key microbloggers such as *@tunkuv* (editor) and *@tjmarx* (filmmaker).

Likewise, we compare the topical configuration of our model *BNTSB.K* to the temporal configuration *BNTSB.KO* where the topical and temporal features are activated. Figure 7.11 presents *MAP* difference of *BNTSB.KS* model from *BNTSB.K* model. Considering all the queries, *BNTSB.KO* model shows an improvement of 17% compared to *BNTSB.K*. We conclude that the temporal distribution is an indicator of tweet relevance.

	BNTS.K	BNTS.KS
MAP	0.1579	**0.1909**

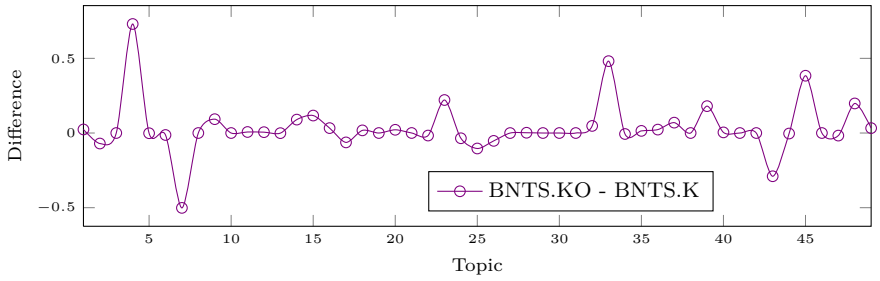


Figure 7.11: BNTS.KO difference from BNTS.K *MAP* per topic

Main improvement of *BNTSB.KO* configuration is observed for topic 4 *"Mexico drug war"* which is in fact a news based topic. Analyzing the related distribution

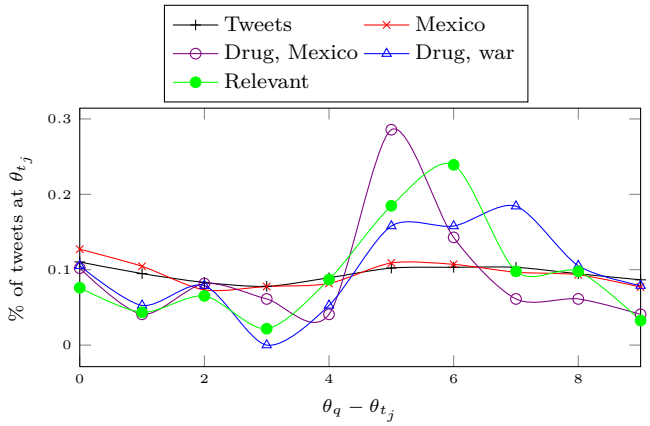


Figure 7.12: Temporal distribution of tweets (topic 4)

of tweets over the time in figure 7.12, we observe that relevant tweets are mainly concentrated in the 5^{th} day before the query. Similar distribution is presented by tweets containing *"Mexico drug"* or *"drug war"* with some decay. Meanwhile, the distribution of all the tweets or tweets containing only the term *"Mexico"* is regular. This confirms our choice to study the temporal distribution per term configuration in *BNTSb* instead of the global distribution of tweets as proposed before in *BNTSi* which may be impacted by commonly used terms.

Conclusion

We proposed in this paper a social model for tweet search that integrates, within a Bayesian network model, the topical relevance of tweets, the social relevance of microbloggers and the temporal relevance of tweet period. In particular, the topical relevance score highlights tweets that present all terms of the query rather that some repeated ones. The social score underlines tweets published by influencer microbloggers. Finally, the temporal score emphasizes tweets published in activity periods of the query topic. Experiments conducted on TREC 2011 Microblog dataset shows that the integration of the different sources of evidence enhances the quality of tweet search.

In future work, we plan to automatically detect the query profile and adjust the score of integrated features according to the sensibility of the query to the social and temporal contexts. In addition, we plan to represent hashtags and URLs entities in the Bayesian network model.

Part III

Conclusion

Chapter 8

Conclusion

8.1 Contributions

The work presented in this thesis is brought into the context of social information retrieval which represents in fact a new research field that bridges information retrieval approaches and social network analysis methods. In particular, we address, on the first hand, the problem of social network integration in the retrieval processes, and on the second hand, the problem of search and access over social networking data. We provide for this aim an overview of the main social information retrieval approaches in the domain of literature access and microblog retrieval. Previous approaches proposed for the first application domain, namely literature access, exploit social networks in order to enhance traditional information retrieval processes. Microblog Retrieval approaches propose however a new methodology for search and access over social networks. The review of the state of the art work has revealed four research questions that deal with (*i*) the social network modeling, (*ii*) the evaluation of social importance of network actors, (*iii*) the relevance modeling in social networks and (*iv*) the integration of social relevance factors. These questions have been addressed by both social information retrieval models proposed in this thesis, in particular for literature access and microblog retrieval. Our contribution can be summarized in three main points: (*1*) the evaluation of social relevance of information entities, (*2*) the identification of key actors in social networks and (*3*) the integration of social relevance factors in the retrieval process. These contributions are sumarized in the following.

1. The relevance of information entities is defined by the importance of related people in the social network. Based on this idea, we propose a social information retrieval model that estimates the relevance of scientific publications in accordance to authors' position the scientific social network as well as the annotators' position in scholarly social bookmarking networks. Authors' social

network is modeled using co-authorships and citation links. Social annotators are represented using friendship and co-tagging networks. Based on these two social networks, we proposed a generic social raking algorithm, called *SoRank*, that allows identifying authoritative experts. In particular, *SoRank* propagates expertise scores of network actors thought network edges in respect to relationships weights. The expertise of actors is computed in this work using classical language models in information retrieval. A social score is attributed to scientific publications as the sum of *SoRank* scores of respective authors and social annotators. Finally, the social score of scientific papers is combined with topical score using linear combination. Experiments on CitaEval evaluation dataset show that our model outperforms traditional information retrieval approaches, scientific impact based approaches and closely related social retrieval approaches with a significant improvement. Results show also the interest of integrating co-authorships and citation links to evaluate the social importance of scientific publications. Finally, considering both authors and annotations social networks lead to comprehensive evaluation of the social relevance of scientific publications and particularly leverage the social relevance of recently published papers that are given enough time to be cited.

2. Key actors in the social network are defined in respect to the social networking activities. In the same purpose of *SoRank* algorithm that helps to indentify authoritative experts in social networks, we proposed in this thesis three link analysis algorithms that identify key actors in microblogging social networks. We focus in the context on the social network topology and we propose to represent microbloggers with a multigprahs where nodes represent microblog authors and edges represent followership, retweets and mentions. These relationships summarize main microblogging activities particularly communication activities and social influence among microbloggers. With regard to the social network structure, we have investigated three types of key microbloggers, namely influencers, leaders and discussers. Influencers are characterized with their ability to spread information though the social network. Leaders are able to engage a large community to realize a common goal. Finally, discussers initiate valuable discussion around interesting topics. In view of that, we introduce three ranking algorithms of key microbloggers inspired by *Pagerank* algorithms, namely *InRank*, *LeadRank* and *Discuss-Rank* that indentify respectively microblog influencers, leaders and discussers. Experimental evaluation conducted over a microblog dataset show that the proposed algorithms outperform microblogging ranking baselines specifically *followers* count commonly used in this context to estimate the importance of microbloggers as well *PageRank* score computed on the followership and retweet networks. Results show also that *LeadRank* shows better precisions compared to other active microblogging ranking algorithms which allows to conclude about the nature of key microbloggers in social networks who exhibit at the same time a high influence and an intensive interaction with their

community. Besides, *InRank* and *DiscussRank* show also intersting results with some notable microbloggers in respect to query topic are ranked in the top results.

3. Relevance definition in social networking environments may integrate different factors. In microblogging context, we propose to integrate the topical relevance, the social relevance, and the temporal relevance. Topical relevance is defined by the textual similarity between the tweet and the query. The social relevance is defined by the position of the microblogger in the social network particularly his influence and leadership as proposed by *InfRank*, *LeadRank* and *DiscussRank* algorithms. In particular, we propose to integrate these relevance factors within a Bayesian network model. Two Bayesian network topologies for microblog search model are introduced in the work including inference Bayesian network and Belief Bayesian networks. Conducted experiments on TREC Microblog Track 2011 and 2012 show that social networks and temporal features may enhance the real-time search within microblogs in particular for socially and temporally queries. Moreover, we conclude that term frequencies are not more informative in comparative to simple presence of query terms in the tweet which is a good of indicator of relevance. Tuning experiments for temporal interval Δt show that best retrieval precision are achieved with Δt value is near to $1day$. Comparing the impact of social importance algorithms, we note slight improvement realized by *leadrank* algorithm compared to *InRank* and *DiscussRank*. Leadership may thus express the social importance of tweets in the context of real-time search. Meanwhile, the performances of these algorithms are limited by evaluation constraints where retweets are presumed irrelevant. Furthermore, we learn from experiments that analyzing the temporal distribution of terms configurations instead of single terms of all the query terms is more useful to detect activity period of the query topic. In fact, this method overcomes the problem of popular terms with consistently high distributions over time. Final results show that our belief Bayesian network model present better results than the Bayesian network model. Although the precisions of our models are quite low compared to first ranked systems in TREC Microblog Track 2011, our belief Bayesian network model overpasses the first system in Microblog Track 2012 with an improvement of 23%.

8.2 Discussions

With the lack of retrieval evaluation standards, the real performances of social retrieval systems remain debatable. Even though current evaluation methods namely TREC standard are more reliable for evaluating Web search and enterprise document search, these methods are topical-biased and do not efficiently evaluate the social search. In evaluation protocol of chapter 5 and 7, domains experts are asked to evaluate the relevance of some documents in response to

predefined queries. Their evaluation is conducted based on textual description of the query. However, this does not reflect real world scenarios of social search where the user is influenced at every step by the social context. We think that the evaluation of social information retrieval systems my conducted using activity logs and search history. Unfortunately, this data is not widely available.

The accuracy of social network data has not been evaluated in this work, for instance, the accuracy of citation links and author name disambiguation in the case of literature access model. The performances of the social retrieval model strongly depends on this social data. On the other hand, microblog dataset is build using a sampling of 1% of twitter stream without any information how Twitter select this data. Regardless of the significance of this sampling, we ignore if there exists other key microbloggers hidden by random sampling unless we analyze the entire social network. Accessing to full data may lead in this case to reliable analysis of social importance of network actors.

Social information retrieval models proposed in this thesis do not address the problem of scalability and latency. One drawback for instance regarding our Bayesian network models is the number of term configuration to evaluate every time. With a query of n terms, 2^n term configuration probabilities must computed. This makes our system particularly slow and not practical for real-time search. Furthermore, the evaluation of the social importance of network actors requires a lot of computing resources. It not envisaged applying our microblogger ranking algorithms on very large networks with millions of nodes. These algorithms may be optimized for large scale network.

8.3 Future Work

With regard to social information retrieval challenges discussed in chapter 1, we have addressed in this thesis the issues of social context evaluation and relevance definition. In future work and from a global perspective of view, we plan to investigate (a) the compositional and the structural divergence of social networking applications as well as (b) data volume and sparsity problems. In particular, the social retrieval models proposed in this thesis will be extended to fit other domain applications. Interestingly, a generic model may ensure in this context a reusable information retrieval system that could be applied of different social networking spaces independently from application purpose or the social network structure. The design of a universal social network model is hence necessary to ensure a retrieval processes at generic level. Regarding data volume and sparsity, we plan to integrate Big Data technologies and techniques and in the core design of our social retrieval model typically by reduced representation of the social network and *HadHoop* algorithms for indexing and social network analysis (Wang et al., 2012; Bruns and Liang, 2012; Manovich, 2011). We expect so more reliable social retrieval model that would easy to implement, practical and efficient in terms of computing time and resource consuming.

At a fine level, we plan to extend our social information retrieval with other entities such as scientific venues and institutions. These entities enhance the social network of scientific publication with a structural layer. Furthermore, we plan we conduct a unified approached for relevance evaluation. In particular, relevance features will be modeled on the social information network using weights on both nodes and edges (Amer-Yahia et al., 2008). A link analysis algorithm will be applied to rank jointly network entities. We expect a multi-entity ranking version of our *SoRank* algorithm that produces several ranking list in accordance with node types namely papers, authors, tags, bookmarking users, scientific venues, etc. A part of this work was already presented in our papers (Soulier et al., 2012b,a, 2013) where we propose to rank jointly authors and papers in heterogeneous bibliographic network.

Our three algorithms *InfRank*, *LeadRank* and *DiscussRank* for identifying, respectively, microblog influencers, leaders and discussers show promising results for ranking microbloggers and tweets. We plan to apply these algorithms for other application purposes typically for top stories identification and social media monitoring. Instead of analyzing the whole social network activities, we propose to focus on a representative sample of key microbloggers to extract trending topics as well the public interest and sentiment of the Internet community. This solution may not be resource-consuming as the case of full network analysis. Moreover, we plan to use our there microblogger ranking algorithms to identify microblog leaders of election campaign in political events. A sentiment analysis component would be integrated in the social network model to conduct this analysis.

We defined in section 4.3.2 several factors of relevance in microblogging context. Our Bayesian network models for tweet search exploit mainly content-based, social-based and temporal-based relevance features. In the future, we plan to integrate more relevance factors within the Bayesian networks namely URLs which show good results in TREC Microblog Track by extending tweets with the content of the respective URL. Location and semantic based features are also envisaged for a better understanding of the tweeting context. In respect of the query profile, we plan to use machine learning methods in to automatically detect the query profile and appropriately weight the integrated features. Face to the computation and the latency issues of our Bayesian network models, we plan to simplify some probabilities with prior probabilities compute only once at the end of the indexing process and permanently stored in the index namely the probability $P(u_f)$ of selecting microblogger u_f and the probability $P(k_i|O_e)$ of observing term k_i in period O_e. Always from a probabilistic point of view, we plan to use the language model framework as a unified model for tweet search. This model may integrate the same features as our Bayesian network model while ensuring reasonable computing time.

Bibliography

Adamic, L. and Adar, E. (2005). How to search a social network. *Social Networks*, 27(3):187–203.

Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 665–674, New York, NY, USA. ACM.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Albakour, M.-D., Macdonald, C., and Ounis, I. (2013). Identifying local events by using microblogs as social sensors. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 173–180, Paris, France, France. Le Centre de Hautes Etudes Internationales d'Informatique Documentaire.

Amati, G., Amodeo, G., Bianchi, M., Celi, A., Nicola, C. D., Flammini, M., Gaibisso, C., Gambosi, G., , and Marcone, G. (2011). Fub, iasi-cnr, univaq at trec 2011. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*.

Amer-Yahia, S., Benedikt, M., and Bohannon, P. (2007). Challenges in searching online communities. *IEEE Data Eng. Bull.*, 30(2):23–31.

Amer-Yahia, S., Galland, A., Stoyanovich, J., and Yu, C. (2008). From del.icio.us to x.qui.site: recommendations in social tagging sites. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 1323–1326, New York, NY, USA. ACM.

Anthonisse, J. (1971). The rush in a graph. *Amsterdam: University of Amsterdam Mathematical Centre*.

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference*

on Empirical Methods in Natural Language Processing, pages 1568–1576. Association for Computational Linguistics.

Armentano, M., Godoy, D., and Amandi, A. (2012). Topology-based recommendation of users in micro-blogging communities. *Journal of Computer Science and Technology*, 27(3):624–634.

Armentano, M. G., Godoy, D., and Amandi, A. (2011). Towards a followee recommender system for information seeking users in twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings*, volume 730, pages 27–38.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison Wesley.

Bai, J., Nie, J.-Y., Cao, G., and Bouchard, H. (2007). Using query contexts in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 15–22, New York, NY, USA. ACM.

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA. ACM.

Balog, K., Azzopardi, L., and de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 43–50, New York, NY, USA. ACM.

Balog, K., de Rijke, M., and Weerkamp, W. (2008). Bloggers as experts: feed distillation using expert retrieval models. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 753–754, New York, NY, USA. ACM.

Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D., and Aly, M. (2008). Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 895–904, New York, NY, USA. ACM.

Bandyopadhyay, A., Ghosh, K., Majumder, P., and Mitra, M. (2012). Query expansion for microblog retrieval. *International Journal of Web Science*, 1(4):368–380.

Bandyopadhyay, A., Mitra, M., and Majumder, P. (2011). Query expansion for microblog retrieval. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. (2007). Optimizing web search using social annotations. In *WWW '07: Proceedings of the* 16^{th}

International Conference on World Wide Web, pages 501–510, New York, NY, USA. ACM.

Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics.

Basu, C., Hirsh, H., Cohen, W., and Nevill-Manning, C. (2001). Technical paper recommendation: A study in combining multiple information sources. *Journal of Artificial Intelligence Research*, 1:231–252.

Baumgartner, H. and Pieters, R. (2003). The structural influence of marketing journals: A citation analysis of the discipline and its subareas over time. *Journal of Marketing*, pages 123–139.

Beauchamp, M. (1965). An improved index of centrality. *Behavioral Science*, 10:161.

Becker, H., Naaman, M., and Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35(12):29–38.

Bellot, P., Chappell, T., Doucet, A., Geva, S., Gurajada, S., Kamps, J., Kazai, G., Koolen, M., Landoni, M., Marx, M., Mishra, A., Moriceau, V., Mothe, J., Preminger, M., Ramírez, G., Sanderson, M., Sanjuan, E., Scholer, F., Schuh, A., Tannier, X., Theobald, M., Trappett, M., Trotman, A., and Wang, Q. (2012). Report on inex 2012. *SIGIR Forum*, 46(2):50–59.

Ben Jabeur, L., Damak, F., Tamine, L., Pinel-Sauvagnat, K., Cabanac, G., and Boughanem, M. (2012a). IRIT at TREC Microblog 2012: Adhoc Task. In Voorhees, E. M. and Buckland, L. P., editors, *Text REtrieval Conference (TREC), Gaithersburg, USA, 07/11/2012-09/11/2012*, page (on line), http://www-nlpir.nist.gov/. National Institute of Standards and Technology (NIST).

Ben Jabeur, L. and Tamine, L. (2010). Vers un modèle de Recherche d'Information Sociale pour l'accès aux ressources bibliographiques (poster). In *Conférence francophone en Recherche d'Information et Applications (CORIA), Sousse, Tunisie*, pages 325–336. Centre de Publications Universitaires.

Ben Jabeur, L., Tamine, L., and Boughanem, M. (2010). A social model for Literature Access: Towards a weighted social network of authors (regular paper). In *International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO), Paris, France*, page (electronic medium). Centre de hautes études internationales d'Informatique Documentaire (C.I.D.). Taux de sélectivité 19

Ben Jabeur, L., Tamine, L., and Boughanem, M. (2011). Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter (regular paper). In *Conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI), Grenoble*, page (en ligne). IMAG.

Ben Jabeur, L., Tamine, L., and Boughanem, M. (2012b). Active microbloggers: Identifying influencers, leaders and discussers in microblogging networks (short paper). In *Symposium on String Processing and Information Retrieval (SPIRE), Cartagena, Colombia, 21/10/2012-25/10/2012*, volume 7608, pages 111–117, http://www.springer.com. Springer Berlin / Heidelberg.

Ben Jabeur, L., Tamine, L., and Boughanem, M. (2012c). Featured tweet search: Modeling time and social influence for microblog retrieval (regular paper). In *IEEE/WIC/ACM International Conference on Web Intelligence, Macau, China, 04/12/2012-07/12/2012*, pages 166–173, http://www.computer.org/portal/web/cscps. IEEE Computer Society - Conference Publishing Services.

Ben Jabeur, L., Tamine, L., and Boughanem, M. (2012d). Intégration des facteurs temps et autorité sociale dans un modèle bayésien de recherche de tweets (regular paper). In *Conférence francophone en Recherche d'Information et Applications (CORIA), Bordeaux, 21/03/2012-23/03/2012*, pages 301–316, http://www.labri.fr. LABRI.

Ben Jabeur, L., Tamine, L., and Boughanem, M. (2012e). Uprising microblogs: A Bayesian network retrieval model for tweet search (regular paper). In *ACM Symposium on Applied Computing (SAC), Riva del Garda (Trento), Italy, 26/03/2012-30/03/2012*, pages 943–948, http://www.acm.org/. ACM.

Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on twitter. In *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.

Bergsma, S., McNamee, P., Bagdouri, M., Fink, C., and Wilson, T. (2012). Language identification for creating language-specific twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 65–74, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bermingham, A. and Smeaton, A. F. (2012). An evaluation of the role of sentiment in second screen microblog search tasks. *Age*, 25(17):18.

Bird, C., Elliott, P., and Griffiths, E. (1996). User interfaces for content-based image retrieval. *IEE Seminar Digests*, 1996(119):8–8.

Bird, C., Gourley, A., Devanbu, P., Gertz, M., and Swaminathan, A. (2006). Mining email social networks. In *Proceedings of the 2006 international workshop on Mining software repositories*, MSR '06, pages 137–143, New York, NY, USA. ACM.

Bischoff, K., Firan, C. S., Nejdl, W., and Paiu, R. (2008). Can all tags be used for search? In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 193–202, New York, NY, USA. ACM.

Bodendorf, F. and Kaiser, C. (2009). Detecting opinion leaders and trends in online social networks. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, pages 65–68. ACM.

Bogers, T. and van den Bosch, A. (2008). Recommending scientific articles using citeulike. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pages 287–290, New York, NY, USA. ACM.

Bollacker, K., Lawrence, S., and Giles, C. (2000). Discovering relevant scientific literature on the web. *Intelligent Systems and their Applications, IEEE*, 15(2):42–47.

Bollen, J., Mao, H., and Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Bollen, J., Rodriquez, M., and Van de Sompel, H. (2006). Journal status. *Scientometrics*, 69(3):669–687.

Bollen, J., Van de Sompel, H., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6):e6022.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, pages 1170–1182.

Bordons, M. and Gómez, I. (2000). Collaboration networks in science. In Garfield, E., Cronin, B., and Atkins, B., editors, *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*, Asis Monograph Series, pages 197–213. Information Today.

Borgman, C. L. (1996). Why are online catalogs still hard to use? *Journal of the American Society for Information Science*, 47(7):493–503.

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10.

Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2):3–10.

Bruns, A. and Liang, Y. E. (2012). Tools and methods for capturing twitter data during natural disasters. *First Monday*, 17(4).

Busch, M., Gade, K., Larson, B., Lok, P., Luckenbill, S., and Lin, J. (2012). Earlybird: Real-time search at twitter. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1360–1369.

Bush, V. (1979). As we may think. *SIGPC Note.*, 1(4):36–44.

Bush, V. and Think, A. (1945). The atlantic monthly. *As We May Think*, 176(1):101–108.

Cabanac, G. and Hartley, J. (2013). Issues of work–life balance among jasist authors and editors. *Journal of the American Society for Information Science and Technology*, pages n/a–n/a.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington DC, USA.

Chemers, M. (1997). *An integrative theory of leadership*. Lawrence Erlbaum Associates.

Chen, C., Li, F., Ooi, B. C., and Wu, S. (2011a). Ti: an efficient indexing mechanism for real-time search on tweets. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, SIGMOD '11, pages 649–660, New York, NY, USA. ACM.

Chen, H.-H., Gou, L., Zhang, X., and Giles, C. L. (2011b). Collabseer: a search engine for collaboration discovery. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 231–240, New York, NY, USA. ACM.

Choi, J. and Croft, W. B. (2012). Temporal models for microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2491–2494. ACM.

Chu, Z., Widjaja, I., and Wang, H. (2012). Detecting social spam campaigns on twitter. In *Applied Cryptography and Network Security*, pages 455–472. Springer.

Cleveland, G. and Dataflow, I. U. (1998). *Digital libraries: definitions, issues and challenges*. IFLA, Universal dataflow and telecommunications core programme.

Cole, J. and Cole, S. (1981). *Social Stratification in Science*. University of Chicago Press.

Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *International AAAI Conference on Weblogs and Social Media*.

Cosijn, E. and Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, 36(4):533 – 550.

Crandall, D., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., and Kleinberg, J. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441.

Craswell, N., Robertson, S., Zaragoza, H., and Taylor, M. (2005). Relevance weighting for query independent evidence. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 416–423, New York, NY, USA. ACM.

Cui, A., Zhang, M., Liu, Y., and Ma, S. (2011). Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *Information Retrieval Technology*, pages 238–249. Springer.

Damak, F., Ben Jabeur, L., Cabanac, G., Pinel-Sauvagnat, K., Tamine, L., and Boughanem, M. (2011). IRIT at TREC Microblog 2011. In Ellen M., V. and Lori P., B., editors, *Text REtrieval Conference (TREC), Gaithersburg, USA, 06/11/2011-09/11/2011*, page (on line), http://www-nlpir.nist.gov/. National Institute of Standards and Technology (NIST).

Damak, F., Pinel-Sauvagnat, K., Boughanem, M., and Cabanac, G. (2013). Effectiveness of state-of-the-art features for microblog search. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 914–919, New York, NY, USA. ACM.

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., and Lee, L. (2009). How opinions are received by online communities: a case study on amazon.com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 141–150, New York, NY, USA. ACM.

Das Sarma, A., Das Sarma, A., Gollapudi, S., and Panigrahy, R. (2010). Ranking mechanisms in twitter-like forums. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 21–30, New York, NY, USA. ACM.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60.

Davies, A. and Ghahramani, Z. (2013). Language-independent bayesian sentiment mining of twitter. In *SNAKDD Workshop 2011 on Social Network Mining and Analysis*.

de Cristo, M. A. P., Calado, P. P., de Lourdes da Silveira, M., Silva, I., Muntz, R., and Ribeiro-Neto, B. (2003). Bayesian belief networks for ir. *International Journal of Approximate Reasoning*, 34(2–3):163 – 179. <ce:title>Soft Computing Applications to Intelligent Information Retrieval on the Internet</ce:title>.

De Solla Price, D. J. (1963). Letter to the editor. *Science*, 139(3555):682.

Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113.

Deerwester, S. (1988). Improving Information Retrieval with Latent Semantic Indexing. In Borgman, C. L. and Pai, E. Y. H., editors, *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, volume 25, Atlanta, Georgia. American Society for Information Science.

Deng, H., Han, J., Lyu, M. R., and King, I. (2012). Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL '12, pages 71–80, New York, NY, USA. ACM.

Deng, H., King, I., and Lyu, M. R. (2008). Formal models for expert finding on dblp bibliography data. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 163–172, Washington, DC, USA. IEEE Computer Society.

Diakopoulos, N. A. and Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM.

Diamond Jr, A. M. (1986). What is a citation worth? *Journal of Human Resources*, pages 200–215.

Ding, Y., Yan, E., Frazho, A., and Caverlee, J. (2009). Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243.

Domingos, P. (2005). Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20(1):80–82.

Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., and Zha, H. (2010). Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 331–340, New York, NY, USA. ACM.

Du, N., Wu, B., Pei, X., Wang, B., and Xu, L. (2007). Community detection in large-scale social networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 16–25, New York, NY, USA. ACM.

Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 295–303, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dumais, S. T. and Nielsen, J. (1992). Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 233–244, New York, NY, USA. ACM.

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., and Vaughan, A. (2010). Omg earthquake! can twitter improve earthquake response? *Seismological Research Letters*, 81(2):246–251.

Earle, P. S., Bowden, D. C., and Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6).

Efron, M. (2010). Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 787–788, New York, NY, USA. ACM.

Efron, M. (2011). Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69:131–152.

Evans, B. M. and Chi, E. H. (2008). Towards a model of understanding social search. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 485–494, New York, NY, USA. ACM.

Eysenbach, G. (2011). Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4).

Fang, H. and Zhai, C. (2007). Probabilistic models for expert finding. In *Proceedings of the 29th European conference on IR research*, ECIR'07, pages 418–430, Berlin, Heidelberg. Springer-Verlag.

Farooq, U., Song, Y., Carroll, J., and Giles, C. (2007). Social bookmarking for scholarly digital libraries. *Internet Computing, IEEE*, 11(6):29–35.

Faust, K. (1997). Centrality in affiliation networks. *Social Networks*, 19(2):157 – 191.

Fawcett, T. (2006). An introduction to {ROC} analysis. *Pattern Recognition Letters*, 27(8):861 – 874. <ce:title>ROC Analysis in Pattern Recognition</ce:title>.

Feng, W. and Wang, J. (2013). Retweet or not?: personalized tweet re-ranking. In *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM '13, pages 577–586, New York, NY, USA. ACM.

Ferguson, P., O'Hare, N., Lanagan, J., Phelan, O., and McCarthy, K. (2012). An investigation of term weighting approaches for microblog retrieval. In *Advances in Information Retrieval*, pages 552–555. Springer.

Fiala, D., Rousselot, F., and Ježek, K. (2008). Pagerank for bibliographic networks. *Scientometrics*, 76(1):135–158.

Fox, E. A. and Sharan, S. (1986). A comparison of two methods for soft boolean operator interpretation in information retrieval. Technical report, Virginia Polytechnic Institute & State University, Blacksburg, VA, USA.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Freeman, L. (1979). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969.

Fu, Y., Xiang, R., Liu, Y., Zhang, M., and Ma, S. (2007). Finding experts using social network analysis. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 77–80, Washington, DC, USA. IEEE Computer Society.

Garcia, R. and Amatriain, X. (2010). Weighted content based methods for recommending connections in online social networks. In *Workshop on Recommender Systems and the Social Web*, pages 68–71.

Garfield, E. (1964). Citation indexes for science. *Readings in information retrieval*, 122(3159):261.

Garg, N. and Weber, I. (2008). Personalized, interactive tag recommendation for flickr. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pages 67–74, New York, NY, USA. ACM.

Gayo-Avello, D. and Brenes, D. (2010). Overcoming spammers in twitter-a tale of five algorithms. In *1st Spanish Conference on Information Retrieval, Madrid, Spain*.

Geisler, G. and Burns, S. (2007). Tagging video: conventions and strategies of the youtube community. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '07, pages 480–480, New York, NY, USA. ACM.

Genc, Y., Sakamoto, Y., and Nickerson, J. V. (2011). Discovering context: classifying tweets through a semantic transform based on wikipedia. In *Proceedings of the 6th international conference on Foundations of augmented cognition: directing the future of adaptive systems*, FAC'11, pages 484–492, Berlin, Heidelberg. Springer-Verlag.

Giles, C. L., Bollacker, K. D., and Lawrence, S. (1998). Citeseer: an automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, DL '98, pages 89–98, New York, NY, USA. ACM.

Girvan, M. and Newman, M. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Givon, S. and Lavrenko, V. (2009). Predicting social-tags for cold start book recommendations. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 333–336, New York, NY, USA. ACM.

Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57(1):27–57.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Gollapalli, S. D., Mitra, P., and Giles, C. L. (2012). Similar researcher search in academic environments. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL '12, pages 167–170, New York, NY, USA. ACM.

Gouws, S., Metzler, D., Cai, C., and Hovy, E. (2011). Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 20–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Goyal, A., Bonchi, F., and Lakshmanan, L. V. (2008). Discovering leaders from community actions. In *CIKM '08*, pages 499–508.

Green, R. M. and Sheppard, J. W. (2013). Comparing frequency-and style-based features for twitter author identification. In *The Twenty-Sixth International FLAIRS Conference*.

Greene, D., Reid, F., Sheridan, G., and Cunningham, P. (2011). Supporting the curation of twitter user lists. *CoRR*, abs/1110.1349.

Grinev, M., Grineva, M., Boldakov, A., Novak, L., Syssoev, A., and Lizorkin, D. (2009). Sifting micro-blogging stream for events of user interest. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 837–837, New York, NY, USA. ACM.

Guan, Z., Wang, C., Bu, J., Chen, C., Yang, K., Cai, D., and He, X. (2010). Document recommendation in social tagging services. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 391–400, New York, NY, USA. ACM.

Guy, I., Zwerdling, N., Ronen, I., Carmel, D., and Uziel, E. (2010). Social media recommendation based on people and tags. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 194–201, New York, NY, USA. ACM.

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E. A. (2003). Automatic document metadata extraction using support vector machines. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, JCDL '03, pages 37–48, Washington, DC, USA. IEEE Computer Society.

Han, Z., Li, X., Yang, M., Qi, H., Li, S., and Zhao, T. (2012). Hit at trec 2012 microblog track. In *Proceeddings of the 21th Text REtrieval Conference (TREC 2010)*.

Hanneman, R. and Riddle, M. (2005). *Introduction to Social Network Methods*. University of California.

Hannon, J., Bennett, M., and Smyth, B. (2010). Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 199–206, New York, NY, USA. ACM.

Harman, D. (1988). Towards interactive query expansion. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '88, pages 321–331, New York, NY, USA. ACM.

Harpale, A., Yang, Y., Gopal, S., He, D., and Yue, Z. (2010). Citedata: a new multi-faceted dataset for evaluating personalized search performance. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 549–558, New York, NY, USA. ACM.

Haustein, S. and Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, 5(3):446–457.

Herbrich, R., Graepel, T., and Obermayer, K. (1999). Large margin rank boundaries for ordinal regression. *Advances in Neural Information Processing Systems*, pages 115–132.

Heymann, P., Koutrika, G., and Garcia-Molina, H. (2008a). Can social bookmarking improve web search? In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 195–206, New York, NY, USA. ACM.

Heymann, P., Ramage, D., and Garcia-Molina, H. (2008b). Social tag prediction. In *SIGIR '08: Proceedings of the* 31^{st} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 531–538, New York, NY, USA. ACM.

Hiemstra, D. (2001). *Using language models for information retrieval*. Taaluitgeverij Neslia Paniculata.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America*, 102(46):16569.

Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In Sure, Y. and Domingue, J., editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426. Springer Berlin Heidelberg.

Hughes, A. L. and Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260.

Hui, P. and Gregory, M. (2010). Quantifying sentiment and influence in blogspaces. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 53–61, New York, NY, USA. ACM.

Hummon, N. and Dereian, P. (1989). Connectivity in a citation network: The development of dna theory. *Social Networks*, 11(1):39–63.

Humphrey, S. M. (1992). Indexing biomedical documents: from thesaural to knowledge-based retrieval systems. *Artificial Intelligence in Medicine*, 4(5):343–371.

Hyland, K. (2003). Self-citation and self-reference: Credibility and promotion in academic publication. *Journal of the American Society for Information Science and Technology*, 54(3):251–259.

Jamali, M. and Ester, M. (2009). Using a trust network to improve top-n recommendation. In *RecSys '09: Proceedings of the* 3^{rd} *ACM Conference on Recommender Systems*, pages 181–188, New York, NY, USA. ACM.

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. (2007). Tag recommendations in folksonomies. In *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD 2007, pages 506–514, Berlin, Heidelberg. Springer-Verlag.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA. ACM.

Jeh, G. and Widom, J. (2002). Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 538–543, New York, NY, USA. ACM.

Jensen, F. (2001). *Bayesian Networks and Decision Graphs*. Statistics for Engineering and Information Science. Springer.

Jiang, J., Hidayah, L., Elsayed, T., and Ramadan, H. (2011). Best of kaust at trec-2011: Building effective search in twitter. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*.

Jin, O., Liu, N. N., Zhao, K., Yu, Y., and Yang, Q. (2011). Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 775–784, New York, NY, USA. ACM.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Jomsri, P., Sanguansintukul, S., and Choochaiwattana, W. (2009). A comparison of search engine using "tag title and abstract" with citeulike—an initial evaluation. In *Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for*, pages 1–5. IEEE.

Jomsri, P., Sanguansintukul, S., and Choochaiwattana, W. (2011). Citerank: combination similarity and static ranking with research paper searching. *Int. J. Internet Technol. Secur. Syst.*, 3(2):161–177.

Jung, J. J. (2012). Online named entity recognition method for microtexts in social networking services: A case study of twitter. *Expert Syst. Appl.*, 39(9):8066–8070.

Kamps, J. (2011). The impact of author ranking in a library catalogue. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, BooksOnline '11, pages 35–40, New York, NY, USA. ACM.

Karimi, S., Yin, J., and Thomas, P. (2012). Searching and filtering tweets: Csiro at the trec 2012 microblog track. In *Proceeddings of the 21th Text REtrieval Conference (TREC 2012)*.

Karimzadehgan, M., White, R., and Richardson, M. (2009). Enhancing expert finding using organizational hierarchies. In Boughanem, M., Berrut, C., Mothe, J., and Soule-Dupuy, C., editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 177–188. Springer Berlin Heidelberg.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kazai, G. and Milic-Frayling, N. (2008). Trust, authority and popularity in social information retrieval. In *Proceedings of the 17th ACM conference on*

Information and knowledge management, CIKM '08, pages 1503–1504, New York, NY, USA. ACM.

Kekäläinen, J. (2005). Binary and graded relevance in ir evaluations-comparison of the effects on ranking of ir systems. *Inf. Process. Manage.*, 41(5):1019–1033.

Khrabrov, A. and Cybenko, G. (2010). Discovering influence in communication networks using dynamic graph analysis. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 288–294, Washington, DC, USA. IEEE Computer Society.

Kinsella, S., Murdock, V., and O'Hare, N. (2011). "i'm eating a sandwich in glasgow": modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 61–68, New York, NY, USA. ACM.

Kirchhoff, L., Stanoevska-Slabeva, K., Nicolai, T., and Fleck, M. (2008). Using social network analysis to enhance information retrieval systems. Technical report, University of St.Gallen - Alexandria Repository (Switzerland).

Kirsch, S., Gnasa, M., and Cremers, A. (2006). Beyond the web: Retrieval in social information spaces. *Advances in Information Retrieval*, pages 84–95.

Kleinberg, J. (2008). The convergence of social and technological networks. *Commun. ACM*, 51(11):66–72.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632.

Knuth, D. (1998). *The art of computer programming: Sorting and searching*. The Art of Computer Programming. Addison-Wesley.

Koolen, M., Kamps, J., and Kazai, G. (2012). Social book search: comparing topical relevance judgements and book suggestions for evaluation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 185–194, New York, NY, USA. ACM.

Koolen, M., Kazai, G., and Craswell, N. (2009). Wikipedia pages as entry points for book search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 44–53, New York, NY, USA. ACM.

Korfiatis, N. T., Poulos, M., and Bokos, G. (2006). Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Information Review*, 30(3):252–262.

Kornai, A. (2008). *Mathematical Linguistics*. Advanced Information and Knowledge Processing. Springer.

Kotov, A., Wang, Y., and Agichtein, E. (2013). Leveraging geographical metadata to improve search over social media. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion,

pages 151–152, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Kotsakis, E. (2002). Structured information retrieval in xml documents. In *Proceedings of the 2002 ACM symposium on Applied computing*, SAC '02, pages 663–667, New York, NY, USA. ACM.

Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Krause, B., Jäschke, R., Hotho, A., and Stumme, G. (2008). Logsonomy - social information retrieval with logdata. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, HT '08, pages 157–166, New York, NY, USA. ACM.

Kumar, N. and Carterette, B. (2013). Time based feedback and query expansion for twitter search. In *Advances in Information Retrieval*, pages 734–737. Springer.

Kumar, R., Novak, J., and Tomkins, A. (2010). Structure and evolution of online social networks. *Link Mining: Models, Algorithms, and Applications*, pages 337–357.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

Kwok, K. L. (1989). A neural network for probabilistic information retrieval. In *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '89, pages 21–30, New York, NY, USA. ACM.

Lampos, V., De Bie, T., and Cristianini, N. (2010). Flu detector-tracking epidemics on twitter. In *Machine Learning and Knowledge Discovery in Databases*, pages 599–602. Springer.

Lanagan, J. and Smeaton, A. F. (2011). Using twitter to detect and tag important events in live sports. *Artificial Intelligence*, pages 542–545.

Lancaster, F. and Fayen, E. (1973). *Information retrieval: on-line*. Information sciences series. Melville Pub. Co.

Larson, M., Soleymani, M., Serdyukov, P., Rudinac, S., Wartena, C., Murdock, V., Friedland, G., Ordelman, R., and Jones, G. J. F. (2011). Automatic tagging and geotagging in video collections and communities. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 51:1–51:8, New York, NY, USA. ACM.

Lawani, S. (1982). On the heterogeneity and classification of author self-citations. *Journal of the American Society for Information Science*, 33(5):281–284.

Lawrence, S., Bollacker, K., and Giles, C. L. (1999a). Indexing and retrieval of scientific literature. In *Proceedings of the eighth international conference on Information and knowledge management*, CIKM '99, pages 139–146, New York, NY, USA. ACM.

Lawrence, S., Lee Giles, C., and Bollacker, K. (1999b). Digital libraries and autonomous citation indexing. *Computer*, 32(6):67–71.

Lee, R., Wakamiya, S., and Sumiya, K. (2011). Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321–349.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9):1303–1319.

Li, J., Tang, J., Zhang, J., Luo, Q., Liu, Y., and Hong, M. (2007). Eos: expertise oriented search using social networks. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1271–1272, New York, NY, USA. ACM.

Li, J. and Willett, P. (2009). Articlerank: a pagerank-based alternative to numbers of citations for analysing citation networks. In *Aslib Proceedings*, pages 605–618. Emerald Group Publishing Limited.

Li, X., Guo, L., and Zhao, Y. E. (2008). Tag-based social interest discovery. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 675–684, New York, NY, USA. ACM.

Li, Y., Zhang, Z., Lv, W., Xie, Q., Lin, Y., Xu, R., Xu, W., Chen, G., and Guo, J. (2011a). Pris at trec2011 micro-blog track. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*.

Li, Y., Zhang, Z., Lv, W., Xie, Q., Lin, Y., Xu, R. X. W., Chen, G., and Guo, J. (2011b). Pris at trec2011 micro-blog track. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*.

Liang, F., Qiang, R., and Yang, J. (2011). Pku_icst at trec 2011 microblog track. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*.

Liang, F., Qiang, R., and Yang, J. (2012). Exploiting real-time information retrieval in the microblogosphere. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL '12, pages 267–276, New York, NY, USA. ACM.

Lim, K. H. and Datta, A. (2012). Following the follower: detecting communities with common interests on twitter. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, HT '12, pages 317–318, New York, NY, USA. ACM.

Lin, Y., Li, Y., Xu, W., and Guo, J. (2012). Microblog retrieval based on term

similarity graph. In *Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on*, pages 1322–1325.

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.

Lindsey, D. (1980). Production and citation measures in the sociology of science: The problem of multiple authorship. *Social Studies of Science*, 10(2):145–162.

Liu, D., Hua, X.-S., Wang, M., and Zhang, H.-J. (2010). Image retagging. In *Proceedings of the international conference on Multimedia*, MM '10, pages 491–500, New York, NY, USA. ACM.

Liu, J., Xuan, Z., Dang, Y., Guo, Q., and Wang, Z. (2007). Weighted network properties of chinese nature science basic research. *Physica A: Statistical Mechanics and its Applications*, 377(1):302–314.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

Liu, X., Bollen, J., Nelson, M. L., and Van de Sompel, H. (2005a). Co-authorship networks in the digital library research community. *Inf. Process. Manage.*, 41(6):1462–1480.

Liu, X., Croft, W. B., and Koll, M. (2005b). Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 315–316, New York, NY, USA. ACM.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4):309–317.

Luo, Z., Osborne, M., and Wang, T. (2012). Opinion retrieval in twitter. In *International AAAI Conference on Weblogs and Social Media*.

Ma, H., Lyu, M. R., and King, I. (2009). Learning to recommend with trust and distrust relationships. In *RecSys '09: Proceedings of the* 3^{rd} *ACM Conference on Recommender Systems*, pages 189–196, New York, NY, USA. ACM.

Ma, N., Guan, J., and Zhao, Y. (2008). Bringing pagerank to the citation analysis. *Inf. Process. Manage.*, 44(2):800–810.

Macdonald, C. and Ounis, I. (2006). Voting for candidates: adapting data fusion techniques for an expert search task. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 387–396, New York, NY, USA. ACM.

Macdonald, C. and Ounis, I. (2008). Voting techniques for expert search. *Knowledge and information systems*, 16(3):259–280.

Manning, C., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, pages 460–75.

Massoudi, K., Tsagkias, M., de Rijke, M., and Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. In *Advances in Information Retrieval*, pages 362–367. Springer.

Mathioudakis, M. and Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD '10, pages 1155–1158, New York, NY, USA. ACM.

McCreadie, R. and Macdonald, C. (2013). Relevance in microblogs: enhancing tweet retrieval using hyperlinked documents. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 189–196, Paris, France, France. Le Centre de Hautes Etudes Internationales d'Informatique Documentaire.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., and McCullough, D. (2012). On building a reusable twitter corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1113–1114, New York, NY, USA. ACM.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., Konstan, J. A., and Riedl, J. (2002). On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, CSCW '02, pages 116–125, New York, NY, USA. ACM.

Meij, E., Weerkamp, W., and de Rijke, M. (2012). Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 563–572, New York, NY, USA. ACM.

Metzler, D. and Cai, C. (2011). Usc/isi at trec 2011: Microblog track. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*.

Mills, A., Chen, R., Lee, J., and Rao, H. R. (2009). Web 2.0 emergency applications: how useful can twitter be for emergency response. *Journal of Information Privacy & Security*, 5(3):3–26.

Mishne, G. (2006). Information access challenges in the blogspace. In *the International Workshop on Intelligent Information Access (IIIA 2006)*. Citeseer.

Mislove, A., Gummadi, K., and Druschel, P. (2006). Exploiting social networks for internet search. In *5th Workshop on Hot Topics in Networks (HotNets06). Citeseer*, page 79.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings*

of the 7th ACM SIGCOMM conference on Internet measurement, IMC '07, pages 29–42, New York, NY, USA. ACM.

Miyanishi, T., Seki, K., and Uehara, K. (2013). Combining recency and topic-dependent temporal variation for microblog search. In *Advances in Information Retrieval*, pages 331–343. Springer.

Müller, J. and Stocker, A. (2011). Enterprise microblogging for advanced knowledge sharing: The referencesbt case study. *j-jucs*, 17(4):532–547.

Mutschke, P. (2003). Mining networks and central entities in digital libraries. a graph theoretic approach applied to co-author networks. *Advances in intelligent data analysis V*, pages 155–166.

Nagmoti, R., Teredesai, A., and De Cock, M. (2010). Ranking approaches for microblog search. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 153–157.

Nakajima, S., Tatemura, J., Hara, Y., Tanaka, K., and Uemura, S. (2006). Identifying agitators as important blogger based on analyzing blog threads. In *APWeb '06*, pages 285–296.

Nallapati, R. (2004). Discriminative models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71. ACM.

Newman, M. E. J. (2001a). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132.

Newman, M. E. J. (2001b). Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64:016131.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

Newman, M. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39 – 54.

Nichols, J., Mahmud, J., and Drews, C. (2012). Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, IUI '12, pages 189–198, New York, NY, USA. ACM.

Nieminen, J. (1974). On the centrality in a graph. *Scandinavian Journal of Psychology*, 15(1):332–336.

O'Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129.

Ota, Y., Maruyama, K., and Terada, M. (2012). Discovery of interesting users in twitter by overlapping propagation paths of retweets. In *Proceedings of the 2012 IEEE/WIC/ACM international Conference on Web Intelligence (WI)*, pages 274–279.

Ounis, I., Macdonald, C., Lin, J., and Soboroff, I. (2011). Overview of the trec-2011 microblog track. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Pal, A. and Counts, S. (2011). Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 45–54, New York, NY, USA. ACM.

Pang, B. and Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Now Publishers.

Parra, D. and Brusilovsky, P. (2009). Collaborative filtering for social tagging systems: an experiment with citeulike. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 237–240, New York, NY, USA. ACM.

Parsons, T. (1949). *The structure of social action: a study in social theory with special reference to a group of recent European writers*. Number vol. 1 in The Structure of Social Action: A Study in Social Theory with Special Reference to a Group of Recent European Writers. Free Press.

Pearl, J. (1985). *Bayesian Networks: A Model of Self-activated Memory for Evidential Reasoning*. Report. UCLA, Computer Science Department.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann Publishers.

Perlman, G. (1993). Information retrieval techniques for hypertext in the semi-structured toolkit. In *Proceedings of the fifth ACM conference on Hypertext*, HYPERTEXT '93, pages 260–267, New York, NY, USA. ACM.

Petkova, D. and Croft, W. B. (2006). Hierarchical language models for expert finding in enterprise corpora. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '06, pages 599–608, Washington, DC, USA. IEEE Computer Society.

Phelan, O., McCarthy, K., and Smyth, B. (2009). Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 385–388, New York, NY, USA. ACM.

Phelan, T. (1999). A compendium of issues for citation analysis. *Scientometrics*, 45(1):117–136.

Pichappan, P. and Sarasvady, S. (2002). The other side of the coin: The intricacies of author self-citations. *Scientometrics*, 54(2):285–290.

Pinheiro, C. (2011). *Social Network Analysis in Telecommunications*. Wiley and SAS Business Series. John Wiley & Sons.

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5):297 – 312.

Pirkola, A. (2001). Morphological typology of languages in IR. *Journal of Documentation*, 57(3):330–348.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

Porter, M. F. (1997). Readings in information retrieval. In Sparck Jones, K. and Willett, P., editors, *Readings in information retrieval*, chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Priem, J. and Costello, K. L. (2010). How and why scholars cite on twitter. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4.

Priem, J. and Hemminger, B. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social web. *First Monday*, 15(7).

Priem, J., Parra, C., Piwowar, H., Groth, P., and Waagmeester, A. (2012a). Uncovering impacts: a case study in using altmetrics tools. In *Workshop on the Semantic Publishing (SePublica 2012) 9 th Extended Semantic Web Conference Hersonissos, Crete, Greece, May 28, 2012*, page 40.

Priem, J., Piwowar, H., and Hemminger, B. (2012b). Altmetrics in the wild: Using social media to explore scholarly impact. *ALM*, 1:3.

Radicchi, F., Fortunato, S., Markines, B., and Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80(5):056103.

Raven, B. H. (1964). *Social Influence and Power*. Defense Technical Information Center.

Ravikumar, S., Balakrishnan, R., and Kambhampati, S. (2012). Ranking tweets considering trust and relevance. In *Proceedings of the Ninth International Workshop on Information Integration on the Web*, IIWeb '12, pages 4:1–4:4, New York, NY, USA. ACM.

Rheingold, H. (2000). *The virtual community: homesteading on the electronic*

frontier. Number n?28 in The Virtual Community: Homesteading on the Electronic Frontier. Mit Press.

Ribeiro, B. A. N. and Muntz, R. (1996). A belief network model for ir. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 253–260, New York, NY, USA. ACM.

Riemer, K. and Richter, A. (2010). Tweet inside: Microblogging in a corporate context. *Proceedings of the 23rd Bled eConference*, pages 1–17.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., and Harabagiu, S. M. (2012). Empatweet: Annotating and detecting emotions on twitter. In *LREC*, pages 3806–3813.

Robertson, S. E. (1991). On term selection for query expansion. *J. Doc.*, 46(4):359–364.

Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 232–241, New York, NY, USA. Springer-Verlag New York, Inc.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1995). Okapi at TREC-3. In *Proceedings of TREC-3 '95*, pages 109–126.

Rodgers, J. L. and Nicewander, A. W. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Roegiest, A. and Cormack, G. V. (2012). University of waterloo: Logistic regression and reciprocal rank fusion at the microblog track. In *Proceeddings of the 21th Text REtrieval Conference (TREC 2012)*.

Romero, D., Galuba, W., Asur, S., and Huberman, B. (2011). Influence and passivity in social media. In Gunopulos, D., Hofmann, T., Malerba, D., and Vazirgiannis, M., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6913 of *Lecture Notes in Computer Science*, pages 18–33. Springer Berlin Heidelberg.

Ross, C., Terras, M., Warwick, C., and Welsh, A. (2010). Pointless babble or enabled backchannel: conference use of twitter by digital humanists. *Digital Humanities*.

Rossi, L. and Magnani, M. (2012). Conversation practices and network structure in twitter. In *International AAAI Conference on Weblogs and Social Media*.

Rowlands, T., Hawking, D., and Sankaranarayana, R. (2010). New-web search with microblog annotations. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1293–1296, New York, NY, USA. ACM.

Rueger, S. (2010). *Multimedia Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA. ACM.

Salton, G. (1991). The smart document retrieval project. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, pages 356–358, New York, NY, USA. ACM.

Salton, G., Fox, E. A., and Wu, H. (1983). Extended boolean information retrieval. *Commun. ACM*, 26(11):1022–1036.

Salton, G. and McGill, M. (1983). *Introduction to modern information retrieval*. McGraw-Hill computer science series. McGraw-Hill.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51, New York, NY, USA. ACM.

Santos-Neto, E., Condon, D., Andrade, N., Iamnitchi, A., and Ripeanu, M. (2009). Individual and social behavior in tagging systems. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09, pages 183–192, New York, NY, USA. ACM.

Sayyadi, H. and Getoor, L. (2009). Future rank: Ranking scientific articles by predicting their future pagerank. In *2009 SIAM International Conference on Data Mining (SDM09)*.

Schafer, J., Frankowski, D., Herlocker, J., and Sen, S. (2007). Collaborative filtering recommender systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 291–324. Springer Berlin Heidelberg.

Schatz, B., Mischo, W., Cole, T., Hardin, J., Bishop, A., and Chen, H. (1996). Federating diverse collections of scientific literature. *Computer*, 29(5):28–36.

Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J. X., and Weikum, G. (2008). Efficient top-k querying over social-tagging networks. In *SIGIR '08: Proceedings of the* 31^{st} *annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 523–530, New York, NY, USA. ACM.

Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. (2010). Folks in folksonomies: social link prediction from shared metadata. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 271–280, New York, NY, USA. ACM.

Schreiber, M. (2007). Self-citation corrections for the hirsch index. *EPL (Europhysics Letters)*, 78(3):30002.

Sen, S., Vig, J., and Riedl, J. (2009). Tagommenders: connecting users to items through tags. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 671–680, New York, NY, USA. ACM.

Serdyukov, P. and Hiemstra, D. (2008). Modeling documents as mixtures of persons for expert finding. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 309–320, Berlin, Heidelberg. Springer-Verlag.

Sigurbjörnsson, B. and van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 327–336, New York, NY, USA. ACM.

Silva, I., Ribeiro-Neto, B., Calado, P., Moura, E., and Ziviani, N. (2000). Link-based and content-based evidential information in a belief network model. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 96–103, New York, NY, USA. ACM.

Singh, L. and Getoor, L. (2007). Increasing the predictive power of affiliation networks. *IEEE Data Eng. Bull.*, 30(2):41–50.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269.

Soboroff, I. and Craswell, N. (2007). Overview of the trec 2006 enterprise track. In *The Fifteenth Text REtrieval Conference (TREC 2006)*, pages 32–51.

Soboroff, I., Ounis, I., Macdonald, C., and Lin, J. (2012). Overview of the trec-2012 microblog track. In *Proceeddings of the 21th Text REtrieval Conference (TREC 2012)*.

Song, X., Chi, Y., Hino, K., and Tseng, B. (2007). Identifying opinion leaders in the blogosphere. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 971–974, New York, NY, USA. ACM.

Soulier, L., Ben Jabeur, L., Tamine, L., and Bahsoun, W. (2012a). BibRank: a language-based model for co-ranking entities in bibliographic networks. In *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), Washington, DC, USA*, pages 61–70. ACM DL.

Soulier, L., Ben Jabeur, L., Tamine, L., and Bahsoun, W. (2012b). Modèle de langue pour l'ordonnancement conjoint d'entités pertinentes dans un réseau d'informations hétérogènes. In *INFormatique des Organisations et Systemes d'Information et de Decision (INFORSID), Montpellier*, page (en ligne). Association INFORSID.

Soulier, L., Ben Jabeur, L., Tamine, L., and Bahsoun, W. (2013). On Ranking Relevant Entities in Heterogeneous Networks Using a Language-Based Model. *Journal of the American Society for Information Science and Technology (JASIST)*, 64(3):500–515.

Sousa, D., Sarmento, L., and Mendes Rodrigues, E. (2010). Characterization of the twitter replies network: are user ties social or topical? In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, pages 63–70, New York, NY, USA. ACM.

Spina, D., Meij, E., de Rijke, M., Oghina, A., Bui, M. T., and Breuss, M. (2012). Identifying entity aspects in microblog posts. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 1089–1090, New York, NY, USA. ACM.

Spink, A., Wolfram, D., Jansen, M. B., and Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American society for information science and technology*, 52(3):226–234.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM.

Strohman, T., Croft, W. B., and Jensen, D. (2007). Recommending citations for academic papers. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 705–706, New York, NY, USA. ACM.

Sun, Y. and Giles, C. L. (2007). Popularity weighted ranking for academic digital libraries. In *Proceedings of the 29th European conference on IR research*, ECIR'07, pages 605–612, Berlin, Heidelberg. Springer-Verlag.

Sun, Y., Li, H., Councill, I. G., Huang, J., Lee, W.-C., and Giles, C. L. (2008). Personalized ranking for digital libraries based on log analysis. In *Proceedings of the 10th ACM workshop on Web information and data management*, WIDM '08, pages 133–140, New York, NY, USA. ACM.

Surowiecki, J. (2005). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few*. Abacus.

Symeonidis, P., Nanopoulos, A., and Manolopoulos, Y. (2008). Tag recommendations based on tensor dimensionality reduction. In *Proceedings of the 2008*

ACM conference on Recommender systems, RecSys '08, pages 43–50, New York, NY, USA. ACM.

Tang, J., Sun, J., Wang, C., and Yang, Z. (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 807–816, New York, NY, USA. ACM.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 990–998, New York, NY, USA. ACM.

Tao, K., Abel, F., Hauff, C., and Houben, G.-J. (2012). What makes a tweet relevant for a topic? In *MSM 2012, Proceedings of the workshop on Making Sense of Microposts (MSM2012), workshop at the 21st World Wide Web Conference 2012*, pages p. 49–56. CEUR Workshop Proceedings at CEUR-WS.org/Vol-838/.

Teevan, J., Ramage, D., and Morris, M. R. (2011). #twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 35–44, New York, NY, USA. ACM.

Thelwall, M., Buckley, K., and Paltoglou, G. (2011). Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

Thelwall, M. and Hasler, L. (2007). Blog search engines. *Online Information Review*, 31(4):467–479.

Ting, I.-H., Chang, P.-S., and Wang, S.-L. (2012). Understanding microblog users for social recommendation based on social networks analysis. *J. UCS*, 18(4):554–576.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, volume 10, pages 178–185.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2011). Election forecasts with twitter how 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4):402–418.

Turtle, H. and Croft, W. B. (1990). Inference networks for document retrieval. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '90, pages 1–24, New York, NY, USA. ACM.

Turtle, H. and Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9(3):187–222.

Uysal, I. and Croft, W. B. (2011). User oriented tweet ranking: a filtering approach to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2261–2264, New York, NY, USA. ACM.

Vickery, G. and Wunsch-Vincent, S. (2007). *Participative Web and User-Created Content: Web 2.0, Wikis and Social Networking*. Organisation for Economic Co-operation and Development.

Vidgen, R., Henneberg, S., and Naudé, P. (2007). What sort of community is the european conference on information systems? a social network analysis 1993–2005. *European Journal of Information Systems*, 16(1):5–19.

Walker, D., Xie, H., Yan, K., and Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010.

Wang, W., Chen, L., Thirunarayan, K., and Sheth, A. P. (2012). Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press.

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., and Smith, M. (2011). Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA. ACM.

Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 261–270, New York, NY, USA. ACM.

Wu, F. and Huberman, B. (2004). Social structure and opinion formation. *arXiv preprint cond-mat/0407252*.

Wu, X., Zhang, L., and Yu, Y. (2006). Exploring social annotations for the semantic web. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 417–426, New York, NY, USA. ACM.

Xue, G.-R., Zeng, H.-J., Chen, Z., Yu, Y., Ma, W.-Y., Xi, W., and Fan, W. (2004). Optimizing web search using web click-through data. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 118–126, New York, NY, USA. ACM.

Yan, E. and Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *J. Am. Soc. Inf. Sci. Technol.*, 60(10):2107–2118.

Yan, E., Ding, Y., and Sugimoto, C. R. (2011). P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*, 62(3):467–477.

Yanbe, Y., Jatowt, A., Nakamura, S., and Tanaka, K. (2007). Can social bookmarking enhance search in the web? In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '07, pages 107–116, New York, NY, USA. ACM.

Yang, J. and Counts, S. (2010). Predicting the speed, scale, and range of information diffusion in twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Yardi, S. and Boyd, D. (2010). Tweeting from the town square: Measuring geographic local networks. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Yarowsky, D. and Florian, R. (1999). Taking the load off the conference chairs: towards a digital paper-routing assistant. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 220–230.

Yi, K. and Chan, L. (2009). Linking folksonomy to library of congress subject headings: an exploratory study. *Journal of Documentation*, 65(6):872–900.

Yin, L., Kretschmer, H., Hanneman, R., and Liu, Z. (2006). Connection and stratification in research collaboration: an analysis of the collnet network. *Information Processing & Management*, 42(6):1599–1613.

Zhang, J., Qu, Y., Cody, J., and Wu, Y. (2010). A case study of micro-blogging in the enterprise: use, value, and related issues. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 123–132, New York, NY, USA. ACM.

Zhang, W., Yu, C., and Meng, W. (2007). Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 831–840, New York, NY, USA. ACM.

Zhang, X., He, B., and Luo, T. (2012). Transductive learning for real-time twitter search. In *International AAAI Conference on Weblogs and Social Media*.

Zhang, Y. (2011). Learning, innovating, and an emerging core of knowledge: A model for the growth of citation networks. *Available at SSRN 1975606*.

Zhang, Z.-K., Zhou, T., and Zhang, Y.-C. (2011). Tag-aware recommender systems: A state-of-the-art survey. *Journal of Computer Science and Technology*, 26:767–777.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *Proceedings of*

the 33rd European conference on Advances in information retrieval, ECIR'11, pages 338–349, Berlin, Heidelberg. Springer-Verlag.

Zhou, D., Orshanskiy, S. A., Zha, H., and Giles, C. L. (2007). Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 739–744, Washington, DC, USA. IEEE Computer Society.

Zhu, B., Gao, J., Han, X., Shi, C., Liu, S., Liu, Y., and Cheng, X. (2012). Ictnet at microblog track trec 2012. In *Proceeddings of the 21th Text REtrieval Conference (TREC 2012)*.

Zikopoulos, I., Eaton, C., and Zikopoulos, P. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. Mcgraw-hill.

Zobel, J., Moffat, A., and Ramamohanarao, K. (1998). Inverted files versus signature files for text indexing. *ACM Trans. Database Syst.*, 23(4):453–490.

Życzkowski, K. (2010). Citation graph, weighted impact factors and performance indices. *Scientometrics*, 85(1):301–315.