A Priori Relevance Based On Quality and Diversity Of Social Signals

Ismail Badache
IRIT Laboratory
University of Toulouse
Toulouse, France
Ismail.Badache@irit.fr

Mohand Boughanem
IRIT Laboratory
University of Toulouse
Toulouse, France
Mohand.Boughanem@irit.fr

ABSTRACT

Social signals (users' actions) associated with web resources (documents) can be considered as an additional information that can play a role to estimate a priori importance of the resource. In this paper, we are particularly interested in: first, showing the impact of signals diversity associated to a resource on information retrieval performance; second, studying the influence of their social networks origin on their quality. We propose to model these social features as prior that we integrate into language model. We evaluated the effectiveness of our approach on IMDb dataset containing 167438 resources and their social signals collected from several social networks. Our experimental results are statistically significant and show the interest of integrating signals diversity in the retrieval process.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Retrieval model, Experimentation

Keywords

Social signals; Signals diversity; Priors; Language models

1. INTRODUCTION

Majority of information retrieval systems (SIR) exploit two classes of sources of evidence to rank documents in response to the user query. The first class, the most exploited, is dependent on the query, it concerns the term statistics such as term frequency, distribution of term in documents. The second class concerns query-independent factors, which measures a kind of quality or importance of the document. Among these factors, the number of incoming links to a document [12], PageRank [5], topical locality [7], the presence of URL [17], document authors [13] and social signals [11].

In this paper we are interested in social signals associated to web resources. Nowadays, web pages include different social network buttons where users can express if they support or recommend content [1]. In this paper, we assume that, in addition to estimating a priori significance based on simply counting signals related to the resource [3], signals diversity associated to a resource can be also an index that indicates a resource interest beyond a social network or a community. The research questions addressed in this paper are the following:

1. How to estimate the signals diversity of a resource?
2. What is the impact of signals diversity on IR system?
3. Is there an influence of the social networks origin on the quality of their signals?

The remainder of this paper is organized as follows. Section 2 reviews some related work. Section 3 describes our social approach. In section 4, we evaluate the effectiveness of our proposed approach and discuss the results. Finally, we conclude the paper and announce some future work.

2. RELATED WORK

While considerable work has been done in the context of social IR there is still a lack of user studies that would analyze the impact of users' actions diversity in a resource from diverse viewpoints. Major existing works [3, 6, 10, 11] focus on how to improve IR effectiveness by exploiting the users' actions and their underlying social network. For instance, Chelaru et al. [6] study the impact of social signals (like, dislike, comment) on the effectiveness of search on YouTube. In [3, 4], the authors show the impact of different time-sensitive signals taken into account individually and grouped as social properties without considering their diversity in the resource.

Our work is distinguished from the state of art. Our approach is based on signals diversity in the resource to improve relevance ranking of conventional text search. We exploit various signals extracted from different social networks. We note that in previous works, diversity has been applied only to the textual content of the document [2] [16].

Our goal is to estimate the significance of the resource by taking into account signals diversity into the resource. These sources of evidence are incorporated into language model that provides a theoretical founded way to take into account the notion of a priori probability of a document.

SIGIR '15, August 09 - 13, 2015, Santiago, Chile
Copyright 2015 ACM ISBN 978-1-4503-3621-5/15/08 ...$15.00.
http://dx.doi.org/10.1145/2766462.2767807.

3. SOCIAL SIGNALS DIVERSITY

Our approach consists in exploiting social signals and their diversity as a priori knowledge to take them into account in retrieval model. We rely on language model to model signals diversity as a prior probability.

3.1 Notation

Social information that we exploit within the framework of our model can be represented by 4-tuple <U, R, A, SN> where *U, R, A, SN* are finite sets of instances: *Users*, *Resources*, *Actions* and *Social Networks*.

Resources. We consider a collection C=$\{D_1, D_2,...D_n\}$ of n documents. Each document (resource) D can be a Web page, video or other type of Web resources. We assume that resource D can be represented both by a set of textual keywords D_w=$\{D_{w_1}, D_{w_2},...D_{w_z}\}$ and a set of social actions A performed on this resource, D_a=$\{D_{a_1}, D_{a_2},...D_{a_m}\}$.

Actions. We assume a set A=$\{a_1, a_2,...a_m\}$ of m actions that users can perform on the resources. These actions represent the relation between users U=$\{u_1, u_2,...u_h\}$ and resources C. For instance, on Facebook, users can perform the following actions on resources: *like*, *share*, *comment*.

Social Properties. We consider a set X=$\{Popularity, Reputation\}$ of 2 social properties that characterize a document D. Each property is quantified by a specific actions group. These properties are modeled as a priori probability.

3.2 Query Likelihood and Document Priors

We exploit language models [15] to estimate the relevance of document to a query. The language modelling approach computes the probability $P(D|Q)$ of a document D being generated by query Q as follows:

$$P(D|Q) \overset{\text{rank}}{=} P(D) \cdot P(Q|D) = P(D) \cdot \prod_{w_i \in Q} P(w_i|D) \quad (1)$$

$P(D)$ is a document prior, i.e. query-independent feature representing the probability of seeing the document. The document prior is useful for representing and incorporating other sources of evidence in the retrieval process. w_i represents words of query Q. Estimating of $P(w_i|D)$ can be performed using different models (Jelineck Mercer, Dirichlet). The main contribution in this paper is how to estimate $P(D)$ by exploiting social signals.

3.3 Estimating Priors

According to our previous approach [3], the priors are estimated by a simply counting of actions performed on the resource. We assume that signals are independent, the general formula is the following:

$$P_x(D) = \prod_{a_i^x \in A} P_x(a_i^x) \quad (2)$$

$P_x(a_i)$ is estimated using maximum-likelihood:

$$P_x(a_i^x) = \frac{Count(a_i^x, D)}{Count(a_\bullet^x, D)} \quad (3)$$

To avoid Zero probability, we smooth $P_x(a_i)$ by collection C using Dirichlet. The formula becomes as follows:

$$P_x(D) = \prod_{a_i^x \in A} \left(\frac{Count(a_i^x, D) + \mu \cdot P(a_i^x|C)}{Count(a_\bullet^x, D) + \mu} \right) \quad (4)$$

$P(a_i^x|C)$ is estimated using maximum-likelihood:

$$P(a_i^x|C) = \frac{Count(a_i^x, C)}{Count(a_\bullet^x, C)} \quad (5)$$

Where:

- $x \in \{P, R\}$ refers to the social property (*popularity* or *reputation*) estimated from a set of specific actions.
- $a_\bullet^x$ the total number of all social signals associated with x property, in document D or in collection C.
- $P_x(D)$ represents the a priori probability of D.
- $Count(a_i^x, D)$ represents number of occurrence of specific action a_i^x performed on a resource. a_i^x designs action a_i related (or used) to measure x property.

We believe that the diversity of signals in a resource is an index that may indicate an interest beyond a social network or a community. Diversity and quantitative distribution of social signals in a resource may be considered as factors of relevance, i.e., a resource dominated by a single signal should be disadvantaged versus a resource with an equitable distribution of the signals. By applying diversity index of Shannon-Wiener [14], the corresponding formula is as follows:

$$Diversity_s(D) = -\sum_{i=1}^{m} P_x(a_i^x) \cdot log(P_x(a_i^x)) \quad (6)$$

Where $P(a_i^x)$ is defined above, and m represents the total number of signals.

The Shannon index is often accompanied by Pielou evenness index [14] :

$$Diversity_s^{Equit}(D) = \frac{Diversity_s(D)}{MAX(Diversity_s(D))} \quad (7)$$

Where:

$$MAX(Diversity_s(D)) = log(m) \quad (8)$$

The a priori probability $P_x(D)$ is estimated using the formula 4 multiplied by the diversity factor. The corresponding formula is as follows:

$$P_x(D) = \left(\prod_{a_i^x \in A} P_x(a_i^x) \right) \cdot Diversity_s^{Equit}(D) \quad (9)$$

4. EXPERIMENTAL EVALUATION

To validate our approach, we conducted a series of experiments on IMDb dataset. We evaluated the impact of signals diversity by combining it with language model.

4.1 Description of Test Dataset

We used a collection IMDb documents provided by INEX[1]. Each document describes a movie, and is represented by a set of metadata, and has been indexed according to keywords extracted from fields [3]. For each document, we collected

[1] https://inex.mmci.uni-saarland.de/tracks/dc/2011/

specific social signals via their corresponding API of 5 social networks listed in table 2. We chose 30 topics with their relevance judgments provided by INEX IMDb 2011[2]. In our study, we focused on the effectiveness of the top 1000 results. Table 1 shows an example of the documents containing social signals. The document URL is given by the following syntax: $http://www.imdb.com/title/\{id\}/$

Table 1: Instances of 2 documents with social signals

Id	Like	Share	Comment	+1
tt1730728	30	11	2	0
tt1922777	12363	11481	20614	238
Id	Bookmark		Tweet	Share(LIn)
tt1730728	0		2	0
tt1922777	12		2522	14

Table 2 presents the social signals that we take into account in order to estimate each social property.

Table 2: Exploited social signals in quantification

Property	Social signal	Network
Popularity	Number of *Comment*	Facebook
	Number of *Tweet*	Twitter
	Number of *Share(LIn)*	LinkedIn
	Number of *Share*	Facebook
Reputation	Number of *Like*	Facebook
	Number of *Mention* +1	Google+
	Number of *Bookmark*	Delicious

4.2 Results and Discussions

We conducted experiments with models based only on content of documents (Lucene Solr model[3] and Hiemstra language model without prior [9]), as well as approaches combining content and social properties as a priori probabilities of document. We note that the best value of $\mu \in [90, 100]$ The results are listed in table 3.

Table 3 summarizes the results of precisions@k for $k \in \{10, 20\}$, nDCG (Normalized Discounted Cumulative Gain) and MAP (Mean Average Precision). In [3, 4], authors have already shown that taking into account the social characteristics without considering diversity improves IR compared to models based only on the topical relevance. In order to check the significance of the results, we performed the Student test [8] and attached * (strong significance against Baseline (A) and (B)) and ** (very strong significance against Baseline (A) and (B)) to the performance number of each row in the table 3 when the p-value < 0.05 and p-value < 0.01 confidence level, respectively.

Before discussing the results we highlight that first we performed a correlation analysis by using the Spearman's correlation coefficient between signals diversity scores in the documents and their relevance. We found that the diversity represents a positive correlation with the relevance (Spearman's rho = 0.19), this result has aroused our interest and encourage us to exploit the diversity in our IR model. Table 3 (D) lists the results obtained by the integration of the signals diversity in the document. We notice that the NDCG and Precisions are generally better than the baseline listed in Table 3 (C) part when diversity is ignored.

We also notice that the both results are much more better than those obtained by topical models (Baseline (A) Without Priors) and Baseline (B) where signals are taken individually. This leads to our first conclusion, if multiple users of different social networks have found that a resource is useful, then it is more likely that other users will find these resources useful too.

Table 3: Results of P@{10, 20}, nDCG and MAP

IR Models	P@10	P@20	nDCG	MAP
(A) Baseline: Without Priors				
Lucene Solr	0.3411	0.3122	0.3919	0.1782
ML.Hiemstra	0.3700	0.3403	0.4325	0.2402
(B) Baseline: Single Priors				
Like	0.3938	0.3620	0.5130	0.2832
Share	0.4061	0.3649	0.5262	0.2905
Comment	0.3857	0.3551	0.5121	0.2813
Tweet	0.3879	0.3512	0.4769	0.2735
+1	0.3826	0.3468	0.5017	0.2704
Bookmark	0.3730	0.3414	0.4621	0.2600
Share (LIn)	0.3739	0.3432	0.4566	0.2515
(C) Baseline: Combination Priors				
TotalFacebook	0.4209	0.4102	0.5681	0.3125
Popularity	0.4316	0.4264	0.5801	0.3221
Reputation	0.4405	0.4272	0.5900	0.3260
All Criteria	0.4408	0.4262	0.5974	0.3300
All Properties	0.4629	0.4509	0.6203	0.3557
(D) With Considering Signals Diversity				
TotalFacebook^{Div}	0.4227*	0.4187*	0.5713*	0.3167*
Popularity^{Div}	0.4403**	0.4288**	0.5983**	0.3320**
Reputation^{Div}	0.4480**	0.4306**	0.6110**	0.3319**
All Criteria^{Div}	0.4463**	0.4318**	0.6174**	0.3325**
All Properties^{Div}	0.4689**	0.4563**	0.6245**	0.3571**

4.3 Quantitative and Qualitative Analysis

To better understand the effect of these social signals on the selection of relevant documents, we analyze their distribution in the all documents returned by 30 topics.

According to Table 4 we notice that the average frequency of the signals in the relevant documents is higher compared to irrelevant documents (ex. the average of *like* is 362 actions in the relevant documents while in irrelevant documents is 61 actions). We also note that the signals coming from Facebook capture the highest number of relevant documents (see Figure 2), knowing that they are many in irrelevant documents but with a much smaller average. This is due to the engagement rate on Facebook and its dynamic growth [1]. Therefore, the distinction between relevant documents and irrelevant documents is much more sensitive to signal frequency, i.e., relevant documents are characterized by a very high number of Facebook signals compared to the irrelevant documents (see Figure 1). The signals *tweet* and +1 come in second position with an average frequency, respectively, 97 and 29 actions in the relevant documents (see Figure 1). The signal of Delicious (Bookmark) is the weakest criterion among these signals, it appears in 429 relevant documents with an average frequency of 13 actions per document only. Concerning the signal of LinkedIn, we note that 95% of its actions of *share* are condensed in 601 relevant documents with an average frequency of 67 actions per document. However, the number of relevant documents captured by Linkedin *share* signal is very low compared to that captured by Facebook signals, this is due to LinkedIn engagement rate which is very low compared to Facebook, but LinkedIn *share* signal represents the most reliable source in terms of trust compared to other social signals. Therefore, we can say that the presence of this signal, whatever its frequency in a document represents a relevance index.

[2] https://inex.mmci.uni-saarland.de/tracks/dc/2011/
[3] http://lucene.apache.org/solr/

Table 4: Statistics on the distribution of the signals in the documents (relevant and irrelevant)

	Relevant documents containing signals			Relevant documents without signals	Irrelevant documents	
	Number of documents	Number of actions	Average	Number of documents	Number of actions	Average
Like	2210	800458	362.1981	555	1678040	61.6133
Share	2357	856009	363.1774	408	1862909	68.4012
Comment	1988	944023	474.8607	777	1901146	69.8052
Tweet	1735	168448	97.0884	1030	330784	12.1455
+1	790	23665	29.9556	1975	49727	1.8258
Bookmark	429	5654	13.1794	2336	20489	0.7523
Share (LIn)	601	**40446**	67.2985	2164	**2341**	0.0859
Total relevant documents : 2765					Total irrelevant documents : 27235	

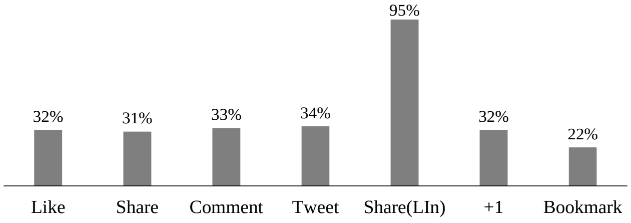

Figure 1: Signals % in the relevant documents

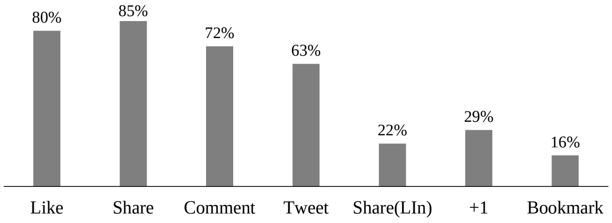

Figure 2: Relevant documents % containing signals

Finally, according to this preliminary statistical study, we observe that each social network has its own specific influence on the quality of its social signals. The quality of signals, provided by Facebook, Twitter, Google+ and Delicious, in a document depend on their frequencies, the more the signals are frequent on the resource, the more its a priori importance increases. However, the LinkedIn signal does not only depend on its frequency in the document because it has in itself a power of mature trust compared to the other signals. This amounts to the maturity of LinkedIn users who are well reputed compared to other social networks users.

5. CONCLUSION

We proposed in this paper a search model based on users' actions associated with a document. We proposed to estimate a social priors of a document by considering signals diversity. The proposed model is based on language model that incorporates this a priori knowledge. Experimental evaluation conducted on IMDb dataset shows that taking into account these social features in a textual model improves the quality of returned search results. Finally, to investigate the influence of social networks on the quality of signals, we developed a statistical study on the distribution of social signals for each social network in both relevant and irrelevant documents.

For future work, we plan to address some limitations of the current study. We plan to integrate other social data into proposed approach. Further experiments on another dataset are also needed. This is even with these simple elements, the first results encourage us to invest more this track.

6. REFERENCES

[1] O. Alonso and V. Kandylas. A study on placement of social buttons in web pages. *arXiv*, 2014.

[2] A. Angel and N. Koudas. Efficient diversity-aware search. In *SIGMOD*, pages 781–792, 2011.

[3] I. Badache and M. Boughanem. Social priors to estimate relevance of a resource. In *IIiX*, 2014.

[4] I. Badache and M. Boughanem. Document priors based on time-sensitive social signals. In *ECIR*, pages 617–622, 2015.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW*, pages 107–117, 1998.

[6] S. Chelaru, C. Orellana-Rodriguez, and I. S. Altingovde. How useful is social feedback for learning to rank youtube videos? *WWW*, pages 1–29, 2013.

[7] Brian D Davison. Topical locality in the web. In *SIGIR*, pages 272–279, 2000.

[8] William Sealy Gosset. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. Originally published under the pseudonym "Student".

[9] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *ECDL*, pages 569–584, 1998.

[10] G. Kazai and N. Milic-Frayling. Effects of social approval votes on search performance. In *ITNG*, pages 1554–1559, 2009.

[11] A. Khodaei and C. Shahabi. Social-textual search and ranking. In *CrowdSearch*, pages 3–8, 2012.

[12] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR*, pages 27–34, 2002.

[13] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM*, pages 387–396, 2006.

[14] EC. Pielou. Shannon's formula as a measure of specific diversity: its use and misuse. *American Naturalist*, pages 463–465, 1966.

[15] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281, 1998.

[16] L. Qin, J. X Yu, and L. Chang. Diversifying top-k results. *VLDB Endowment*, 5(11):1124–1135, 2012.

[17] T. Westerveld, W. Kraaij, and D. Hiemstra. Retrieving web pages using content, links, urls and anchors. 2002.