Social Priors to Estimate Relevance of a Resource

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

In this paper we propose an approach that exploits social data associated with a Web resource to measure its a priori relevance. We show how these interaction traces left by the users on the resources, which are in the form of social signals as the number of *like* and *share*, can be exploited to quantify social properties such as popularity and reputation. We propose to model these properties as a priori probability 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 social properties in a search model to enhance the information retrieval.

Categories and Subject Descriptors

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

General Terms

Retrieval model, Experimentation

Keywords

Social signals, Priors, Language models, Feature evaluation

1. INTRODUCTION

Information retrieval systems (SIR) aim at searching relevant documents in response to user's need. These documents are returned in decreasing order of relevance. Most information retrieval models use the term statistics, such as term frequency, distribution of term in documents. In addition to term statistics, IR models are often extended with further source of evidence often query-independent evidence such as, the number of incoming links to a document [15], its PageRank [5] and the type of its associated URL [15].

One of the important sources which can also be used to measure the a priori interest of Web resources is social data (signals) associated with Web resource resulting from user interaction with this resource [18]. These interactions representing *annotations*, *comments* or *votes*, produce useful and interesting social information that characterizes a resource in terms of popularity and reputation [1, 2, 3]. Major search engines integrate social signals (e.g. Google, Bing). Searchmetrics[1] showed that it exists a high correlation between social signals and the rankings provided by search engines such Google. This paper describes an approach that exploits social signals generated by users on the resources to estimate a priori relevance of a resource. This a priori knowledge is combined with topical relevance modeled by a language modeling (LM) approach. The research questions addressed in this paper are the following:

1. How to translate social signals into social properties?
2. What are the most useful signals and properties to evaluate a priori relevance (importance) of a resource?
3. What theoretical model to combine a priori relevance of resource with its topical relevance?
4. What is the impact of social properties on IR system performance?
5. What are the most favoured signals and properties while using attribute selection algorithms? and what are the most correlated with documents relevance?

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

In this section, we report related work exploiting social signals to measure a priori relevance of a resource.

Some approaches focus on how to improve information retrieval (IR) effectiveness by exploiting users' actions and their underlying social network. *Chelaru et al.*[6] study the impact of social signals (*like*, *dislike*, *comment*, etc.) on the effectiveness of search on YouTube[2]. They show that, although the basic criteria using the similarity of query with video title and annotations are effective for video search, social criteria are also useful and improve the ranking of

[1] www.searchmetrics.com/en/services/ranking-factors-2013/
[2] https://www.youtube.com/

IIix'14, August 26 - 29, Regensburg, Germany

http://dx.doi.org/10.1145/2637002.2637016.

search results for 48% queries. They used "greedy feature selection algorithm" and six learning algorithms. Our approach exploits the same principle, but contrary to the above approach, we do not use learning models, and we exploit more signals from multiple social networks. *Karweg et al.* [12] propose an approach combining topical score and social score based on two factors: first, user engagement intensity quantifies the effort a user has made during an interaction with document, measured by the number of *clicks*, number of *votes*, number of *records* and *recommendation*, secondly, trust degree measured from social graph for each user according to his popularity, using PageRank algorithm. They have found that social results are available for most queries and usually lead to more satisfying results. Similarly, *Khodaei and Shahabi* [14] propose a ranking approach based on several social factors including relationships between document owners and querying user, the importance of each user and user action (*playcount*: number of times a user listens to a track on lastfm[3]) performed on Web documents. They have conducted an extensive experiments set on "lastfm" dataset. They showed a significant improvement for socio-textual ranking compared to the textual only and social only approaches. Compared to the last two approaches, our approach does not take into account the user aspect, we do not use a linear combination, and we exploit more signals from multiple social networks, and we combine them as properties such as popularity and reputation. On Twitter[4], *Hong et al.* [10] use *retweets* as a measure of popularity of tweet and apply machine learning techniques to predict how often new messages will be retweeted. They exploited different features, the content of messages, temporal information, metadata of messages and users, and the user's social graph. Our approach exploits several signals extracted from various social networks.

Finally, there are other studies initiated by Microsoft Bing researchers [17, 23] that show the usefulness of different social contents generated by the network of user friends on Facebook[5]. *Kazai and Milic-Frayling* [13] incorporate different types of social approval *votes* for book IR using external resources that refer to books in the corpus, such as lists from libraries and publishers and lists of bestsellers and award winning books. They define a set of features to compute the social static rank and train the neural network to integrate it with full-text search. They observe the effect of individual features and show that the representations of the general consumer appeal tend to be more effective. They find that social approval *votes* can improve a BM25F baseline that indexes both full-text and MARC[6] records. *Pantel et al.* [19] study the leverage of social annotation on the quality of search results. They observe that the social annotations can benefit Web search in two aspects: first, the annotations are usually good summaries of corresponding web pages; second, the annotations indicate the interest and popularity of web pages. They learned that social aspects are most influential in perceived utility, in particular affinity (degree of closeness), expertise and interest valence (*share, like* and *dislike*). They further established these close social connections and experts in the search topic provide the most utility, whereas distant friends and friends that show no positive or negative interest valence provide the least utility, by a factor of over 50%. These approaches exploit internal social signals to the experimental dataset, whereas our approach exploits external signals from multiple social networks and combine them as properities.

Such previous works our goal aims at exploiting social signals to improve accuracy and relevance of convention textual Web search. We exploit various signals extracted from different social networks. In addition, instead of considering social features separately as done in the previous works, we propose to combine them to measure specific social properties, namely the popularity and the reputation of a resource. We also evaluate the impact of the freshness of the signal in the performance. In our work, we use language models that provide a theoretical founded way to take into account the notion of a priori probabilities of a document.

3. SOCIAL IR APPROACH

Our approach consists of exploiting social signals as a priori knowledge to define social properties to take into account in retrieval model. We rely on language model to combine topical relevance of a given resource to a query and its importance modeled as a prior probability.

3.1 Notation

Social information that we exploit within the framework of our model can be represented by 5-tuple <U, R, A, T, SN> where *U, R, A, T, SN* are finite sets of instances: *Users*, *Resources*, *Actions*, *Times* and *Social networks*.

3.1.1 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=$\{w_1, w_2,...w_z\}$ and a set of social actions A performed on this resource, D_a=$\{a_1, a_2,...a_m\}$.

3.1.2 Actions

We consider a set A=$\{a_1, a_2,...a_m\}$ of m actions (signals) 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*.

3.1.3 Social Properties

We consider a set X=$\{Popularity, Reputation, Freshness\}$ of 3 social properties that characterize a document D. Each property is quantified by a specific actions group. These properties are modeled as a priori probability of a resource.

3.1.4 Time

The time represents the history of each social action, let T_{a_i}=$\{t_{1,a_i}, t_{2,a_i},..t_{k,a_i}\}$ a set of k moments (date) at which action a_i was produced. A moment t represents the datetime for each action a of the same type.

3.2 Query Likelihood and Document Priors

We exploit language models to measure the relevance of document to a query. The language modelling approach computes the probability $P(D|Q)$ of a document D being

[3]http://www.lastfm.fr/
[4]https://twitter.com/
[5]https://www.facebook.com/
[6]http://www.loc.gov/marc/

generated by a query Q as follows using the Bayes theorem [22, 9]:

$$score(Q, D) = P(D|Q) = \frac{P(D) \cdot P(Q|D)}{P(Q)} \quad (1)$$

$P(D)$ is a document prior probability i.e. query-independent feature representing the probability of seeing the document. Typically, this probability is assumed to be the same for any document, hence the document prior is taken to be uniform [25]. Alternatively, the document prior is useful for representing and incorporating other sources of information to the retrieval process. $P(Q)$ can be ignored since it does not depend on the documents and, therefore, does not affect the ranking of documents. $P(Q|D)$ can be represented by a document-based unigram language model:

$$P(D|Q) \stackrel{\text{rank}}{=} P(D) \cdot P(Q|D) = P(D) \cdot \prod_{w_i \in Q} P(w_i|D) \quad (2)$$

Where w_i represents words of the query Q. Estimating of $P(w_i|D)$ can be performed using different models (Jelineck Mercer, Dirichlet)[25]. The main contribution in this paper is how to estimate $P(D)$ using social signals.

To estimate a priori probabilities of resource $P(D)$, we have several options. Either we consider each action individually, in this case, we got as much probabilities than actions. Each $P(D)$ measures the impact of a given action relatively to the other actions in the document or in a set of documents. Either we cumulate or we compute the joint effect of all observed actions in a document, each document is associated one probability. Or, that is what we propose, we believe that social signals indicate a certain user engagement and have different significancy. A *like* type signal does not have the same impact than a *comment* signal.

Therefore, we propose to combine signals according to the properties that might be represented. We then estimate the document prior according to the actions group of this property. we studied three types of social properties of a resource: a) popularity, b) reputation and c) freshness. The first two properties are quantified based on the number of occurrences of one or more specific actions. While freshness is measured from the dates of actions on resource.

3.2.1 Popularity P

It is a social phenomenon which indicates which is the most known among the public. Thanks to the influence of peers, target resources progress quickly in terms of its pervasive in the society. Therefore, Web resource popularity can be estimated according to the rate of sharing this resource between the users through social actions. We assume that the popularity comes from users' activities on social networks, i.e. a resource is said popular if it was *shared* and *commented* by several users in several social networks, to the point where it becomes very known to the general public.

3.2.2 Reputation R

The resource popularity does not reflect its good or bad reputation. Resource reputation is an opinion on this resource, we believe that the estimation of this property can be calculated based on social activities that have positive meaning such as Facebook *like* or *marking* resource as favorites on Delicious[7]. Indeed, resource reputation depends on degree of users' appreciation on social networks.

[7]https://delicious.com/

3.2.3 Freshness F

The freshness is an important relevance factor exploited by several search engines. The information freshness is often measured in relation to its publication date, but we cannot say that information is necessarily obsolete because it was published two years ago. Taking an example of a resource published in September 2001, carrying an information about the attack on "World Trade Center", in 2013, the same resource was discussed in social networks through different social signals. We assume that a resource is fresh if recent social data were associated with it. For that purpose, we define freshness as follows: *"a date of each social action (e.g., date of comment, date of share) performed on a resource on social networks can be exploited to measure the recency of these social actions, hence freshness of information".*

3.3 Estimating Priors

We distinguish two main priors (popularity and reputation). They are estimated by a simply counting the number of specific actions performed on the resource. The general formula is the following:

$$P_x(D) = \prod_{a_i^x \in A} P_x(a_i^x) \quad (3)$$

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

$$P_x(a_i^x) = \frac{Count(a_i^x, D)}{Count(a_\bullet^x, D)} \quad (4)$$

To avoid Zero probability, we smooth $P_x(a_i^x)$ by collection C using Dirichlet [25]. The formula becomes as follows:

$$P_x(D) = \prod_{a_i^x \in A} \left(\frac{Count(a_i^x, D) + \mu \cdot P(a_i^x|C)}{Count(a_\bullet^x, D) + \mu} \right) \quad (5)$$

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

$$P(a_i^x|C) = \frac{Count(a_i^x, C)}{Count(a_\bullet^x, C)} \quad (6)$$

Where:

- $x \in \{P, R\}$ refers to the social property (Popularity or Reputation) estimated from a set of specific actions.
- $P_x(D)$ 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. $a_\bullet^x$ is the total number of social signals associated to x property, in documents D or in collection C.

In addition to simple counting of social actions, we propose to consider the time associated with signal. We assume that the resource associated with fresh (recent) signals should be promoted comparing to those associated with old signals. Each time a given signal appears, it is associated with its occurrence time. Therefore, instead of counting each occurrence of a given signal, we bias this counting, noted $Count_B$, by the date of the occurrence of the signal. The corresponding formula is as follows:

$$Count_B(t_{j,a_i^x}, D) = \sum_{j=1}^{k} f_F(t_{j,a_i^x}, D) = \sum_{j=1}^{k} \exp\left(-\frac{\| t_{current} - t_{j,a_i^x} \|^2}{2\sigma^2}\right) \quad (7)$$

Where:

- $f_F(t_{j,a_i^x}, D)$ represents freshness function, estimated by using Gaussian Kernel [21], it calculates a distance between current time $t_{current}$ and action time t_{j,a_i^x}.
- Gaussian Kernel parameter $\sigma \in R_+$.

Notice that if the time is not considered $f_F(t_{j,a_i^x}, D) = 1$ $\forall t_{j,a_i^x}$, formula 9 will be identical to formula 5, $Count()$ is replaced by $Count_B()$.

3.4 Combining Priors

In our case, we have various sources of social information that influences the a priori probability of relevance. This probability is calculated by combining two main social properties (popularity and reputation). Regarding [20] the problem can be formalized as follows:

$$P_{P\oplus R}(D) = P_P(D) \cdot P_R(D) \quad (8)$$

Where:

- $P_P(D)$, $P_R(D)$ define a priori probabilities relative to popularity P and reputation R that include freshness function.
- $P_{P\oplus R}(D)$ defines the probability of priors combination.

4. EXPERIMENTAL EVALUATION

To evaluate our approach, we conducted a series of experiments on IMDb dataset. We compared our approach which combines document prior with a language model, to the baseline formed by only a language model. In the first instance, the baseline of our evaluation is a retrieval process without the use of any document priors. Our goals in these experiments are:

- first, to evaluate whether social signals, taken from different social networks improve the search.
- second to evaluate the impact of each signal taken separately and grouped to represent a certain property.
- and finally to measure the impact of the freshness.

Moreover, we apply several feature selection algorithms and compute the Spearman's rank correction [4] between the priors and the relevance of documents. Therefore, observing a high correlation would support our hypothesis related to using social signals to enhance a search.

4.1 Description of Test Dataset

We used a collection IMDb documents provided by INEX. Each document describes a movie, and is represented by a set of metadata, and has been indexed according to keywords extracted from fields with status *indexed* in table 1. For each document, we collected specific social signals via their corresponding API of 5 social networks listed in table 2 and table 4. We have put them in the UGC (User Generated Content) field. This field has not been indexed. The nature of these social signals is a counting of each specific social actions on the resource. We chose 30 topics from the set of INEX IMDb topics[8] (see table 3). To obtain relevance judgments, we use Qrels provided by INEX IMDb 2011. In our study, we focused on the effectiveness of the top 1000 results.

Table 1: List of the different document fields

Field	Description	Status
ID	Identifying the film (document)	-
Title	Film's title	Indexed
Year	Year of the film release	Indexed
Rated	Film classification by content type	-
Released	Date of making the film	Indexed
Runtime	Length of the film	Indexed
Genre	Film genre (Action, Drama, etc.)	Indexed
Director	Director of the film project	Indexed
Writer	Writers and writers of the film	Indexed
Actors	Main actors of the film	Indexed
Plot	Text summary of the film	Indexed
Poster	URL of the link poster	-
url	URL of the Web source document	-
UGC	Social data recovered	-

Table 2 shows an example of the documents social data. The document URL is given by the following syntax: *http* : *//www.imdb.com/title/{id}/*

Table 2: Instances of 2 documents with social data

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

4.2 Quantifying Social Properties

Table 4 presents the social signals that we considered in order to estimate each social property (P, R, F).

Specific social signals have been associated with each property depending on their nature and meaning. In table 4, we note that the social signals that quantify reputation carry positive opinions, e.g. *bookmark* resource link by user on Delicious means that this resource has been added to his favorites list. Concerning *like* and *mention* +1, user clicks on these buttons to indicate that he has enjoyed the resource content. So the presence of these social data in the resource increases the degree of resource reputation. The same applies to popularity which is estimated by the exploited social signals that allow us to know the position of this resource

[8] www.inex.otago.ac.nz/

Table 3: Instances of INEX IMDb test topics

Topic	Description	Narrative
action biker	search for all action movies with bikers in it.	As i like action movies, specially if bikers are in it, i like to get a list of all these movies.
ancient Rome era	find the movies about the era of ancient Rome.	I am interested in the movies about era of ancient Rome. I am looking for movies talking stories in the era of ancient Rome.
true story drugs +addiction -dealer	find movies about drugs (drug addiction but not drug dealers) that are based on a true story.	I am working with teens and I want to show them a movie about drugs that is based on a true story. A relevant movie is any true story based movie about drug use and addiction. Movies about drug dealers are not relevant. I would like to see as much information as possible about the movie in order to decide whether the movie is appropriate or not.

Table 4: Exploited social signals in quantification

Property	c_i	Social signal	Network
Popularity	c_1	Number of *Comment*	Facebook
	c_2	Number of *Tweet*	Twitter
	c_3	Number of *Share(LIn)*	LinkedIn
	c_4	Number of *Share*	Facebook
Reputation	c_5	Number of *Like*	Facebook
	c_6	Number of *Mention* +1	Google+
	c_7	Number of *Bookmark*	Delicious
Freshness	c_8	Date of last *Share*	Facebook
	c_9	Date of last *Comment*	Facebook

on the Web, in terms of trend and propagation. Finally, as the date of the different actions are not available except the last date of Facebook actions (*comment* and *share*). We used formula 5 biased only by the last date of *comment* and *share*. The revised formula is as follows:

$$P_x(D) = P_x(D) \cdot f_F(t_{l,c_4}, D) \cdot f_F(t_{l,c_1}, D) \qquad (9)$$

where $f_F(t_{l,\{c_1,c_4\}}, D)$ is calculated like in formula 9, with t_l representing the date of last action (*comment* and *share*).

4.3 Baselines

We used Lucene Solr engine[9] for indexing and retrieval. In all our experiments, during indexing, standard stopwords removal and Porter's stemming algorithm are applied. We used default settings of Lucene solr and Language Model (Hiemstra) [9] as baseline models.

- **Lucene Solr** is a popular search engine developed by the Apache Software Foundation that employs the well-known vector space model of information retrieval and tf-idf term weighting.
- **Language Model (LM.Hiemstra)** denotes a classical IR matching model that computes the query-entity similarity by a smoothed language model, namely the Hiemstra model. The language model is used in our model to compute the content-based score.

[9]http://lucene.apache.org/solr/

4.4 Result and Discussion

We conducted experiments with models based only on content of documents (Lucene Solr model 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 $\mu \in [90, 100]$.

Table 5: Results of P@{10, 20}, nDCG and MAP

IR Models	P@10	P@20	nDCG	MAP
Lucene Solr	0.3411	0.3122	0.3919	0.1782
ML.Hiemstra	0.3700	0.3403	0.4325	0.2402
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
Combination Priors				
$(c_4 + c_1)$	0.4202*	0.4118*	0.5677*	0.3122*
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**
With Integration of Freshness				
Share^F	0.4148*	0.3681*	0.5472*	0.2970*
Comment^F	0.3861*	0.3601*	0.5207*	0.2844*
$(c_4^F + c_1^F)$	0.4310**	0.4220**	0.5806**	0.3174**
Popularity^F	0.4488**	0.4307**	0.6070**	0.3339**
All Criteria^F	0.4475**	0.4282**	0.6053**	0.3348**
All Properties^F	**0.4762****	**0.4701****	**0.6294****	**0.3600****

Table 5 summarizes the results of precisions [16] for 10 and 20 top documents, nDCG (Normalized Discounted Cumulative Gain) [11] and MAP (Mean Average Precision) [16]. We evaluated different configurations, by taking into account social signals and properties individually with and without *freshness* factor, as well their overall combination. In order to check the significance of the results, we performed the Student test [7] and attached * (strong significance against

Hiemstra) and ** (very strong significance against Hiemstra) to the performance number of each row in the table 5 when the p-value < 0.05 and p-value < 0.01 confidence level, respectively.

We observe in all cases, that social signals or social properties significantly improve the results compared to the two baseline models. The results show also that integrating *freshness* provides better results than when it is ignored. The overall combination of social properties (with and without *freshness*) provides the good results, knowing that the best results are obtained with integrating the *freshness*. According to Student test, majority of the results show a statistically significant improvement.

More precisely, we notice that considering social signals, even individually improves the search compared to the baseline. The improvements are lower than when they are combined. We notice also their impact is different. Some signals such as *bookmark* (Delicious) and *share* (LinkedIn) have no impact (statistically not significant), however, signals such as Facebook *share* and Facebook *like* improve the performance in all considered points.

The prior based on properties improve significantly the results compared to the baseline and to individual signal. The *reputation* provides the better results compared to *popularity* and the combination of *share* with *comment*. We recall that we considered this later combination because c_4 and c_1 are the only signals associated with the *freshness* in our experiments. One of the reasons of these results is that the signals that quantify *reputation* may be seen as expressing the engagement of a user who provides his explicit endorsement. For example, the resources having more positive signals (*like* and *mention +1*) are more trustworthy than the ones that do not possess these social signals. If multiple users have found that the resource is useful, then it is more likely that other users will find these resources useful too. The social signals that quantify the *popularity* do not represent approval votes, as for example the *comment* can be positive or negative, but they represent trend factors and a measure of information propagation. Therefore, a popular information always arouses the interest of the user. The other point is that the combination of social signals from various social networks offers more realistic collective judgment of the resource notoriety and builds trust and credibility. An interesting resulting comes from the way all criteria are combined, we notice that combination of properties (named *All Properties*) leads to better results (+5% P@10) than combination of all criteria (named *All Criteria*). This further shows that it is indeed more efficient to apply a smoothing on the social properties than on the social criteria.

We investigate the retrieval performance attainable by integrating the *freshness*, whose priors are estimated based on the simply counting of signal biased by the date of last signal (in our case, date of last *comment* and *share*). Table 5 (With Integration of Freshness) shows that the nDCG and precisions are in general slightly better than the nDCG and precision scores where *freshness* is ignored, but remain very comparable. We notice that the combination ($c_4^F + c_1^F$) with considering the *freshness* leads to better results (+2.57% P@10 and +2.27% nDCG) than ($c_4 + c_1$) when *freshness* is ignored. We also observe that, with *freshness*, the overall combination of the social properties (named *All Properties*^{F}) brings better results (+6.41% P@10 and +4% nDCG) compared to the overall combination of criteria (named *All Criteria*^{F}). Accordingly, these results confirm our hypothesis, that the resources associated with fresh (recent) signals seem to be promoted comparing to those associated with old signals.

Finally, the best results are obtained by *All Properties*^{F} run with rates improvements +50% MAP and +45% nDCG, compared to language model baseline *ML.Hiemstra*. Therefore, grouping the social signals according to their meaning and nature, where some signals are related to *popularity* and others related to *reputation*, with considering the *freshness* factor is the most effective solution to enhance a search.

4.5 Feature Selection Algorithms Study

In order to better understand the real impact of the different signals we conducted a feature selection study, relied on algorithms for selecting attributes to determine the best social signals to exploit in the retrieval model. Feature selection Algorithms [8] aim to identify and eliminate as many irrelevant and redundant information as possible. We used Weka[10] for this experiment. It is a powerful open-source Java-based learning tool that brings together a large number of learning machines and algorithms for selecting attributes.

We proceeded as follows: the top 1000 resources for 30 topics were extracted using Lucene Solr model. Then, the scores of all criteria (social signals) are calculated for each resource. We identify relevant resources and irrelevant according to the qrels. The set of resources obtained contains 30000 instances composed of 2765 relevant resources and 27235 irrelevant resources. We observed that this collection has an unbalanced relevance classes distribution. This occurs when there are many more elements in one class than in the other class of a training collection. In this case, a classifier usually tends to predict samples from the majority class and completely ignore the minority class [24]. For this reason, we applied an approach to subsampling (reducing the number of samples that have the majority class) to generate a balanced collection composed of 2765 relevant resources and 2765 irrelevant resources that were randomly selected. Finally, we applied the attribute selection algorithm on the whole set.

Table 6 shows the social signals selected through attribute selection algorithms. We use ranking methods to rank the selected criteria. The "folds number" in the table indicates how many times the social signal has been selected in the cross-validation task. We note that the signals *share (LIn)* and *bookmark* are seldom favored by selection algorithms. The *mention +1* is moderately favored but it is selected by each algorithm, which indicates its importance even if it is not the best. Thus, the Facebook signals *like*, *share*, *comment* and *tweet* were the highest ranked and often validated over the 4 or 5 iterations of cross-validation.

By comparing these results with the results listed in table 5, social signals (*share (LIn)* and *bookmark*) that provide the lowest results (statistically not significant) are the least favored by the selection algorithms. Also, social signals (*like* and *share*) that provide the best results are highly favored and well ranked by the various selection algorithms.

4.6 Ranking Correlation Analysis

In order to analyze social signals and determine if there is a link (dependence / independence) between them and the

[10]http://www.cs.waikato.ac.nz/ml

Table 6: Selected social signals with attribute selection algorithms

			Popularity				Reputation		
Algorithm	Metric	LS	*Comment*	*Tweet*	*Share(LIn)*	*Share*	*Like*	+1	*Bookmark*
CfsSubsetEval	[folds number]	5	5	5	-	5	5	2	-
WrapperSubsetEval	[folds number]	5	1	1	1	4	5	3	2
ConsistencySubsetEval	[folds number]	5	5	5	5	5	5	5	4
FilteredSubsetEval	[folds number]	5	5	5	-	5	5	2	-
	Average	5	4	4	1.5	4.75	5	3	1.5
ChiSquaredAttributeEval	[rank]	1	4	5	7	2	3	6	8
FilteredAttributeEval	[rank]	1	4	5	7	2	3	6	8
GainRatioAttributeEval	[rank]	1	2	5	8	3	4	6	7
InfoGainAttributeEval	[rank]	1	4	5	7	2	3	6	8
OneRAttributeEval	[rank]	1	3	5	7	4	2	6	8
ReliefFAttributeEval	[rank]	1	4	8	6	2	3	5	7
SVMAttributeEval	[rank]	1	6	7	3	2	5	4	8
SymetricalUncertEval	[rank]	1	2	5	7	3	4	6	8
	Average	1	3.62	5.62	6.5	2.5	3.37	5.62	7.75

document relevance, thus that between them in pairs, we conducted a correlation study. Our goals are as follows:

- first, determine social signals and social properties (with and without *freshness*) correlated with the relevance.
- second, determine the redundant signals, and those that have a same effect on the retrieval improvement.

4.6.1 Correlation Between Signals and Relevance

According to a June 2013 study from Searchmetrics[11], among 22 ranking factors identified, social signals account for 5 of the 6 most highly correlated with Google search results. In addition, BrightEdge[12] survey released in 2012, 84% of search marketers say social signals such as *like*, *tweet*, and *mention +1* will be either more important (53%) or much more important (31%) to their SEO (Search Engine Optimization) compared to 2011.

We analyzed the ranking correlation performed through the Spearman's Rho (r_s) rank correlation coefficient [4], that measures the agreement between each social signal and documents relevance.

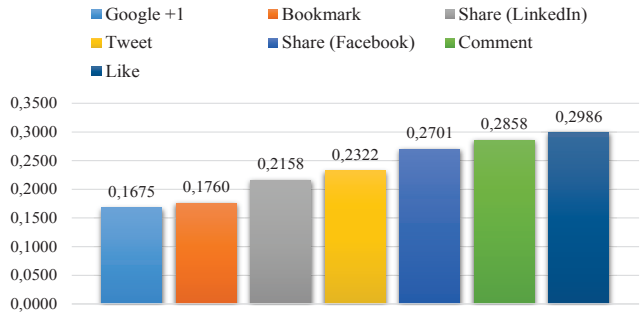

Figure 1: *Rho* Correlations of social signals

Figure 1 shows the values of correlations between ranges social signals with respect to documents relevance. The study shows that *Facebook like* (0.29) has the highest correlation, followed by number of *Fecebook comment* (0.28). Other high-ranking factors include *Facebook share* (0.27) and *tweet* (0.23).

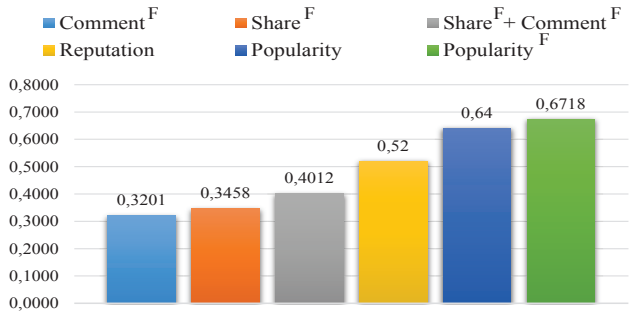

Figure 2: *Rho* correlations of social signals (integrating freshness) and social properties (with and without integrating freshness)

Figure 2 shows the values of correlations between ranges social signals and social properties (with and without integrating *freshness* factor) with respect to documents relevance. The study shows that $Popularity^F$ (0.6718) has the highest correlation compared to all other social priors.

This observation supports the motivation of our proposed combination method, which assumes that among the social signals examined, there are those related to *popularity* and others related to *reputation*. We note that the integration of the *freshness* has improved the rate of correlation with relevant documents. This study justifies the results obtained above (see table 5) and confirms that the temporal aspect of social action also contributes to the improvement of the retrieval performance.

Finally, the ranking correlation analysis shows that all social signals are positively correlated with relevance. Our study confirms the interest of social signals exploited: Well positioned resources have a high number of *like*, *share* and specific resources stand out in the top search results with a very high mass of social data. On the one hand, this means that the activity on social networks continues to increase, on the other hand, it means that the frequently *liked* or *shared* content is increasingly correlated with good ranking of relevance.

[11] www.searchmetrics.com/en/services/ranking-factors-2013/
[12] www.marketingcharts.com/direct/social-signals-increasingly/

Table 7: Spearman's *Rho* correlation values for the social signals pairs

Social Signals	*Like*	*Share*	*Comment*	*Share(LIn)*	*Tweet*	*Bookmark*	*Google + 1*
Like	1						
Share	0.61	1					
Comment	0.31	0.26	1				
Share(LIn)	0.35	0.41	0.40	1			
Tweet	0.32	0.28	0.39	**0.77**	1		
Bookmark	0.34	0.48	0.51	0.31	**0.76**	1	
Google + 1	0.34	0.61	0.40	0.32	0.30	**0.71**	1

4.6.2 Pair-wise Correlation Between Social Signals

To examine the linear relationship for each pair of social signals, we compute the pair-wise overlap between the features by averaging the similarity of their top-1000 rankings over all queries. Atypical method for measuring the similarity of two ranked lists is using the Spearman's *Rho* metric. The more *Rho* is close to 1 (in absolute value), the more the relation is strong and vice-versa.

In table 7, we provide the Spearman's *Rho* scores that are normalized to [0,1] range where 0 means completely different rankings and 1 means equal rankings. The lower diagonal of the table presents the correlation of social signals based on the rankings for all queries. We find that, the top-1000 rankings provided by the social signals pairs (*tweet, share(LIn)*), (*bookmark, Tweet*) and (*mention +1, bookmark*) are highly correlated, i.e., the similarity scores of these pairs are higher than 0.70 (see table 7). These correlations between social signals imply some redundancy, at least for the purposes of ranking. These observations justify and confirm the results obtained by using feature selection approach to filter and rank such redundant social signals. In this study, social signals: *bookmark*, *share (LIn)* are the less important criteria followed by *mention +1*.

Finally, this is a preliminary correlation study, we are well aware that further reflection to better address these issues is needed.

5. CONCLUSION

We proposed in this paper a search model of Web resources based on social properties. These properties, which are considered as a priori probabilities, were defined through social signals. 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 properties in a textual model improves the quality of returned search results. We used feature selection algorithms to identify the best social signals for this task of information retrieval. By analyzing ranking correlations, we note that all social signals present a positive correlation. Meanwhile, this correlation agreement justifies the significant improvement for our social approach.

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] I. Badache. Ri sociale : intégration de propriétés sociales dans un modèle de recherche. In *10th French Information Retrieval COnference*, CORIA'13, pages 305–310, 2013.

[2] I. Badache and M. Boughanem. Exploitation des signaux sociaux pour estimer la pertinence a priori d'une ressource. In *11th French Information Retrieval COnference*, CORIA'14, pages 163–178, 2014.

[3] I. Badache and M. Boughanem. Harnessing Social Signals to Enhance a Search. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WIC'14, Poland, 2014. IEEE Computer Society.

[4] S. D. Bolboaca and L. Jantschi. Pearson versus spearman, kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 5(9):179–200, 2006.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW'98, pages 107–117, Amsterdam, The Netherlands, 1998. Elsevier Science Publishers B. V.

[6] S. V. Chelaru, C. Orellana-Rodriguez, and I. S. Altingovde. Can social features help learning to rank youtube videos? In *Proceedings of the 13th International Conference on Web Information Systems Engineering*, WISE'12, pages 552–566, Berlin, Heidelberg, 2012. Springer-Verlag.

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

[8] Mark A. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. on Knowl. and Data Eng.*, 15(6):1437–1447, November 2003.

[9] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Research and Advanced Technology for Digital Libraries*, volume 1513 of *Lecture Notes in Computer Science*, pages 569–584. Springer Berlin Heidelberg, 1998.

[10] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 57–58, NY, USA, 2011. ACM.

[11] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[12] B. Karweg, C. Huetter, and K. Böhm. Evolving social search based on bookmarks and status messages from

social networks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1825–1834, New York, NY, USA, 2011. ACM.

[13] G. Kazai and N. Milic-Frayling. Effects of social approval votes on search performance. In *Information Technology: New Generations, ITNG '09. Sixth International Conference on*, pages 1554–1559, 2009.

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

[15] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 27–34, New York, NY, USA, 2002. ACM.

[16] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[17] M. R. Morris and J. Teevan. Exploring the complementary roles of social networks and search engines. In *Human-Computer Interaction Consortium Workshop*, pages 1–10, Asilomar, CA, 2012.

[18] A. Oghina, M. Breuss, M. Tsagkias, and M. de Rijke. Predicting imdb movie ratings using social media. In *Advances in information retrieval*, pages 503–507. Springer, 2012.

[19] P. Pantel, M. Gamon, O. Alonso, and K. Haas. Social annotations: Utility and prediction modeling. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 285–294, New York, NY, USA, 2012. ACM.

[20] J. Peng, C. Macdonald, B. He, and I. Ounis. Combination of document priors in web information retrieval. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, RIAO '07, pages 596–611, Paris, France, France, 2007. Le Centre De Hautes Etudes Internationales d'Informatique Documentaire.

[21] Jeff M. Phillips and S. Venkatasubramanian. A gentle introduction to the kernel distance. *CoRR*, abs/1103.1625, 2011.

[22] Jay M. Ponte and W. Bruce Croft. 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, 1998. ACM.

[23] M. Raza. A new level of social search: Discovering the user's opinion before he make one. In *Microsoft Research*, pages 1–6, 2011.

[24] S-J. Yen and Y-S. Lee. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In De-Shuang Huang, Kang Li, and GeorgeWilliam Irwin, editors, *Intelligent Control and Automation*, volume 344 of *Lecture Notes in Control and Information Sciences*, pages 731–740. Springer Berlin Heidelberg, 2006.

[25] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004.