Document Priors Based On Time-Sensitive Social Signals

Ismail Badache and Mohand Boughanem

IRIT - Paul Sabatier University, Toulouse, France
{Badache,Boughanem}@irit.fr

Abstract. Relevance estimation of a Web resource (document) can benefit from using social signals. In this paper, we propose a language model document prior exploiting temporal characteristics of social signals. We assume that a priori significance of a document depends on the date of users actions (social signals) and on the publication date (first occurrence) of the document. Particularly, rather than estimating the priors by simply counting signals related to the document, we bias this counting by taking into account the dates of the resource and the action. We evaluate our approach on IMDb dataset containing 167438 resources and their social data collected from several social networks. The experiments show the interest of temporally-aware signals at capturing relevant resources.

Keywords: Social Information Retrieval, Social Signals, Signal Time, Resource Publication Date, Social Ranking, Language Models.

1 Introduction

Web search engines are expected to return relevant search results for a query. Classic notions of relevance focus on textual relevance. Recently, majority of search engines include social signals (e.g. +1, like) as non-textual features to relevance. However, in the existing works signals are considered time-independent. They are taken into account by only counting the signal frequency on a resource.

In this paper, we hypothesise that signals are time-dependent, the date when the user action has happened is important to distinguish between recent and old signals. Therefore, we assume tha the recency of signals may indicate some recent interests to the resource, which may improve the a priori relevance of document. Secondly, number of signals of a resource depends on the resource age. Generally, an old resource may have much more signals than a recent one.

We introduce the time-aware social approach that incorporates temporal characteristics of users' actions as prior in the retrieval model. Precisely, instead of assuming uniform document priors in this retrieval model, we assign document priors based on the signals associated to that document biased by both the creation date of the signals and the age of the document. Research questions addressed in this paper are the following:

1. How to take into account signals and their date to estimate the priors?
2. What is the impact of temporally-aware signals on IR system performance?

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 617–622, 2015.

The remainder of this paper is organized as follows. Section 2 reviews some related work. Section 3 presents details of our approach. In section 4, we describe our experiments. Finally, we conclude the paper and announce some future work.

2 Related Work

While considerable work has been done in the context of temporal query classification there is still lack of user studies that would analyze users' actions in temporal search from diverse viewpoints. Major existing works [1, 2] focus on how to improve IR effectiveness by exploiting users' actions and their underlying social network. For instance, Chelaru et al. [2] study the impact of social signals (like, dislike, comment) on the effectiveness of search on YouTube. Badache and Boughanem [1] show the impact of different signals individually and grouped.

The works that are most related to our approach include [3, 4], which attempt to improve ranking in Web search. Inagaki et al. [3] propose a set of temporal click features, called ClickBuzz, to improve machine learning recency ranking by favoring URLs that have been of recent interest for users. Khodaei and Alonso [4] propose incorporating time as aspect when investigating social search. They categorized user social interests into five classes: recent, ongoing, seasonal, past and random, and then analyzed Twitter and Facebook data on users activities.

Our work has a similar motivation as those previous efforts, i.e., harnessing any temporal features around a resource to improve relevance ranking of conventional text search. However, our approach is based on novel characteristics which are incorporated into language model. Our goal is to estimate the significance of a resource by taking into account the signal recency and the age of the resource.

3 Time-Aware Social Signals

Our approach focuses on the temporal dimension of users' actions. We rely on language model to model temporally-aware signals as a prior probability.

3.1 Preliminaries and Context

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*.

Resources. We consider a collection C=$\{D_1, D_2,...D_n\}$ of n documents. Each document 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 D, D_a=$\{a_1, a_2,...a_m\}$.

Actions. We consider a set A=$\{a_1, a_2,...a_m\}$ of m actions (signals) that users can perform on resources. These actions (e.g. *like*, *share*, *comment* on Facebook) represent the relation between users U=$\{u_1, u_2,...u_h\}$ and resources C.

Time. Time T represents two types of temporal dimensions:

1. 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 (datetime format) at which action a_i was produced, noted t_{k,a_i}.
2. Age of resource, let T_d=$\{t_{D_1}, t_{D_2},..t_{D_n}\}$ a set of n dates (datetime format) at which each resource D was published, noted t_D.

3.2 Query Likelihood and Document Priors

We exploit language models [5] 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:

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

$P(D)$ 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 to 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) [5]. The main contribution in this paper is how to estimate $P(D)$ by exploiting social signals.

3.3 Estimating Time-Aware Priors

According to our previous approach [1], 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:

$$P(D) = \prod_{a_i \in A} P(a_i) \tag{2}$$

$$P(a_i) \text{ is estimated using maximum-likelihood: } P(a_i) = \frac{Count(a_i, D)}{Count(a_\bullet, D)} \tag{3}$$

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

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

$$P(a_i|C) \text{ is estimated using maximum-likelihood: } P(a_i|C) = \frac{Count(a_i, C)}{Count(a_\bullet, C)} \tag{5}$$

Where: $P(D)$ represents the a priori probability of D. $Count(a_i, D)$ represents number of occurrence of action a_i on resource D. $a_\bullet$ is the total number of social signals in document D or in collection C.

We assume that this simple counting of signals may boost old resources compared to recent ones, because resources with long life in the Web has much more chance to get more signals than recent ones. In addition, we assume that resources that have recent signals are more likely to interest user. We propose to consider the dates associated with a signal and the creation of a resource. To estimate priors, we distinguish two ways to handle it:

a. By considering time of signal: we assume that a 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 the counting, noted $Count_{t_a}$, by the date of the occurrence of the signal.

$$Count_{t_a}(t_{j,a_i}, D) = \sum_{j=1}^{k} f(t_{j,a_i}, D) = \sum_{j=1}^{k} \exp\left(-\frac{\| t_{current} - t_{j,a_i} \|^2}{2\sigma^2}\right) \quad (6)$$

Where: $f(t_{j,a_i}, D)$ represents signal-time function, we use Gaussian Kernel [6] to estimate a distance between current time $t_{current}$ and t_{j,a_i} with $\sigma \in R_+$.

The prior $P(D)$ is estimated using formula 4 but by replacing $Count()$ by $Count_{t_a}()$. Notice that if the signal time is not considered $f(t_{j,a_i}, D) = 1\ \forall t_{j,a_i}$.

b. By considering the age of resource: the resource publication date plays an important role on the social life of this resource, i.e. an old resource has a greater chance to have a large number of interactions compared to a recently published resource. So to cope with this issue we propose to normalize the distribution of signals associated with a resource through resource publication date. We divide the number of signals by the current lifespan of the resource.

$$Count_{t_D}(a_i, D) = \frac{Count(a_i, D)}{Age(D)} = \frac{Count(a_i, D)}{\exp\left(-\frac{\|t_{current} - t_D\|^2}{2\sigma^2}\right)} \quad (7)$$

The prior $P(D)$ is estimated using formula 4 but by replacing $Count()$ by $Count_{t_D}()$ for document and $Count_{t_C}()$ for collection.

4 Experimental Evaluation

To evaluate our approach, we conducted a series of experiments on IMDb dataset. The baseline is a retrieval process without using document priors. Our main goal in these experiments is to evaluate the impact of temporally-aware signals on IR.

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

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

according to keywords extracted from fields [1]. For each document, we collected specific social data via their corresponding API of 5 social networks listed in table 1. The nature of these social signals is a counting of each social actions on the resource. 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 a document with their social data.

Table 1. Instance of document with social data

		Facebook			Google+	Delicious	Twitter	LinkedIn
Film Title	Id	Like	Share	Comment	+1	Bookmark	Tweet	Share
Sinister	*tt1922777*	14763	13881	22914	341	12	2859	14
		Facebook						
Film Title	Id	Last Share		Last Comment		Publication Date		
Sinister	*tt1922777*	2014-09-29T02:49:01		2014-09-28T00:41:01		2011-05-07T19:00:57		

Unfortunately, the date of the different actions are not available except the last date of Facebook actions (*comment* and *share*). Therefore, we represent results using formula 6 biased only by the last date of *comment* and *share*.

4.2 Result and Discussion

We conducted experiments with models based only on documents (Lucene Solr model and Hiemstra language model without prior [7]), as well as approaches combining textual content and social features with temporal aspects as prior of document. We note that the best value of $\mu \in [90, 100]$.

Tables 2 summarizes the results of precision@k for $k \in \{10, 20\}$, nDCG (Normalized Discounted Cumulative Gain) and MAP. We evaluated different configurations, by taking into account social actions, actions time (labeled signal^{T_a}) and resource age (labeled signal^{T_D}). We have already shown that exploiting time-independent signals as prior improve search. In order to check the significance of the results, we performed the Student test and attached * (significance against baselines) to the performance number of each row in the table 2 when the p-value is 0.05 confidence level, compared to the corresponding baselines results.

First, we investigate the retrieval performance attainable by considering the *action time*, in our case date of last *comment* and *share*. Table 2 (With Considering Action Time) shows that the nDCG and precisions are in general slightly better than the nDCG and precision scores where *action time* is ignored, but remain very comparable. Second, we investigate the retrieval performance attainable by considering the *publication date of resource*.Table 2 (With Considering Age of Resource) shows that the nDCG and precisions are in general better than the nDCG and precision scores where *publication date* is ignored (Without Considering Time). Finally, the best results are obtained by (All Criteria^{T_D}) run with considering the *publication date*. Therefore, *publication date* factor is the most effective temporal aspect to enhance a search. Concerning date of signals, we did not really evaluated the real impact of the proposal because of the lack

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

of suitable data (dates of different actions). We exploited only the date of the last action which is not enough to draw effective conclusion.

Table 2. Results of P@k for $k \in \{10, 20\}$, nDCG and MAP

IR Models	P@10	P@20	nDCG	MAP	IR Models	P@10	P@20	nDCG	MAP
Baselines: Without Priors					**With Considering Action Time T_a**				
Lucene Solr	0.3411	0.3122	0.3919	0.1782	Share^{T_a}	0.4148*	0.3681*	0.5472*	0.2970*
ML.Hiemstra	0.3700	0.3403	0.4325	0.2402	Comment^{T_a}	0.3861*	0.3601*	0.5207*	0.2844*
Baselines: Without Considering Time [1]					**With Considering Publication Date T_D**				
Like	0.3938	0.3620	0.5130	0.2832	Like^{T_D}	0.4091*	0.3620*	0.5308*	0.2907*
Share	0.4061	0.3649	0.5262	0.2905	Share^{T_D}	0.4177*	0.3721*	0.5544*	0.2989*
Comment	0.3857	0.3551	0.5121	0.2813	Comment^{T_D}	0.3912*	0.3683*	0.5285*	0.2874*
Tweet	0.3879	0.3512	0.4769	0.2735	Tweet^{T_D}	0.3918*	0.3579*	0.4903*	0.2779*
+1	0.3826	0.3468	0.5017	0.2704	+1^{T_D}	0.3900	0.3511	0.5246	0.2748
Bookmark	0.3730	0.3414	0.4621	0.2600	Bookmark^{T_D}	0.3732	0.3427	0.4671	0.2618
Share (LIn)	0.3739	0.3432	0.4566	0.2515	Share^{T_D} (LIn)	0.3762	0.3449	0.4606	0.2542
All Criteria	0.4408	0.4262	0.5974	0.3300	All Criteria^{T_D}	0.4484*	0.4305*	0.6200*	0.3366*

5 Conclusion

In this paper, we studied the impact of time related to users' actions and resource on IR. We proposed to estimate a social priors of a document by considering the time of the action and the publication date of the resource. Experiments conducted on IMDb dataset show that taking into account social features and temporal aspects in a textual model improves the quality of returned search results. The main contribution of this work is to show that time of user's action and the ratio of signals are fruitful for IR systems. An important issue that we did not address is the exploitation of times associated for each action. Unfortunately, currently social networks APIs do not allow extraction of these informations. For future work, we plan to estimate the impact of signals diversity with respect of their ages. Further experiments on another dataset are also needed.

References

[1] Badache, I., Boughanem, M.: Social priors to estimate relevance of a resource. In: IIiX Conference. IIiX 2014, pp. 106–114. ACM, NY (2014)
[2] Chelaru, S., Orellana-Rodriguez, C., Altingovde, I.S.: How useful is social feedback for learning to rank youtube videos? In: World Wide Web, pp. 1–29 (2013)
[3] Inagaki, Y., Sadagopan, N., Dupret, G., Dong, A., Liao, C., Chang, Y., Zheng, Z.: Session based click features for recency ranking. In: AAAI Press (2010)
[4] Khodaei, A., Alonso, O.: Temporally-aware signals for social search. In: SIGIR 2012 Workshop on Time-aware Information Access (2012)
[5] Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR Conference, pp. 275–281. ACM, NY (1998)
[6] Phillips, J.M., Venkatasubramanian, S.: A gentle introduction to the kernel distance. CoRR abs/1103.1625 (2011)
[7] Hiemstra, D.: A linguistically motivated probabilistic model of information retrieval. In: Nikolaou, C., Stephanidis, C. (eds.) ECDL 1998. LNCS, vol. 1513, pp. 569–584. Springer, Heidelberg (1998)