Priors Based On Time-Sensitive Social Signals

Ismail Badache, Mohand Boughanem

Priors Based On Time-Sensitive Social Signals

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

1. Introduction

- 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.
- **Hypothesis 1:** signals are time-dependent, the date when the user action has happened is important to distinguish between recent and old signals. Therefore, the recency of signals may indicate some recent interests to the resource, which may improve the a priori relevance of document.
- **Hypothesis 2:** the number of signals of a resource depends on the resource age, an old resource may have much more signals than a recent one.

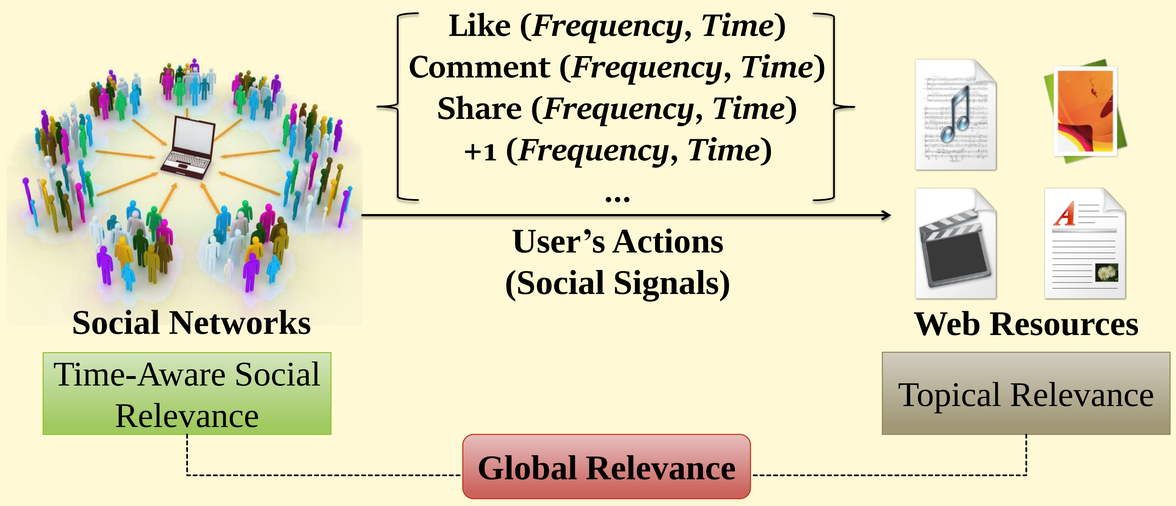


Fig. Global presentation of our approach

- **Research questions 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?

2. Time-Independent Social Signals

- We exploit language model to estimate the relevance of document D to a query Q.

$$P(D|Q) =^{Rank} P(D) \cdot P(Q|D) = P(D) \cdot \prod_{w_i \epsilon Q} P(w_i|Q) \quad (1)$$

- $\boldsymbol{P(D)}$ is a document prior. w_i represents words of query Q. Estimating $P(w_i|Q)$ can be performed using different models.
- The priors are estimated by a counting of actions a_i performed on D.

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

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

- We smooth $P(a_i)$ by collection C using *Dirichlet*:

$$\boldsymbol{P(D)} = \prod_{a_i \in A} \left(\frac{Count(a_i, D) + \mu \cdot P(a_i|C)}{Count(a_\bullet, D) + \mu} \right) \quad (3)$$

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

Where $\boldsymbol{P(D)}$ 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.

3. Time-Aware Social Signals

- We propose to consider the date associated with a signal and the age of a resource. To estimate priors, we distinguish two ways to handle it:

- **Proposal 1: Time of Signal**

Resource associated with fresh signals are more likely to interest user and should be promoted comparing to those associated with old signals. 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.

$$Count_{t_a}(t_{j,a_i}, D) = \sum_{j=1}^{k} f(t_{j,a_i}, D) = \sum_{j=1}^{k} exp\left(-\frac{\| t_{current} - t_{j,a_i} \|^2}{2\sigma^2} \right) \quad (4)$$

Where $f(t_{j,a_i}, D)$ represents signal-time function, we use Gaussian Kernel to estimate a distance between current time $t_{current}$ and t_{j,a_i} with $\sigma \in \Re^+$. k represents the number of moments (*datetime*) at which action a_i was produced.

- $\boldsymbol{P(D)}$ is estimated using formula 3 but by replacing $Count()$ by $Count_{t_a}()$. Notice that if the signal time is not considered $f(t_{j,a_i}, D) = 1 \ \forall t_{j,a_i}$

- **Proposal 2: Age of Resource**

The 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. 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.

$$Count_{t_D}(a_i, D) = \frac{Count(a_i, D)}{Age(D)} = \frac{Count(a_i, D)}{\exp\left(-\frac{\| t_{current} - t_D \|^2}{2\sigma^2} \right)} \quad (5)$$

- The prior $\boldsymbol{P(D)}$ is estimated using formula 3 but by replacing $Count()$ by $Count_{t_D}()$ for document and $Count_{t_C}$ for collection.

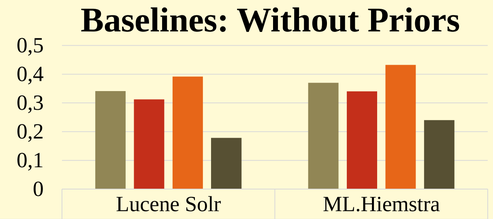
4. Experimental Evaluation

- **Dataset**

- INEX IMDb Dataset.
- 30 INEX IMDb Topics and their relevance judgments.
- Social data from 5 social networks: *Like*, *Share* and *Comment* (Facebook); *Tweet* (Twitter); *+1* (Google+); *Bookmark* (Delicious); *Share* (LinkedIn).

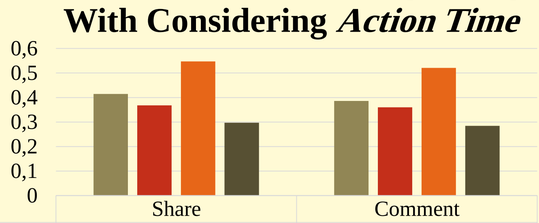
- **Results**

[Rank : 4]
Baselines: Without Priors



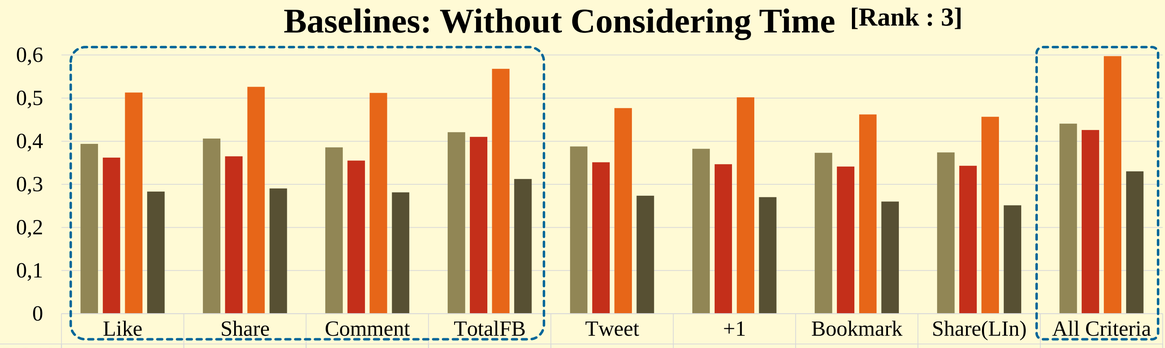
	Lucene Solr	ML.Hiemstra
P@10	0,3411	0,37
P@20	0,3122	0,3403
nDCG	0,3919	0,4325
MAP	0,1782	0,2402

[Rank : 2]
With Considering *Action Time*



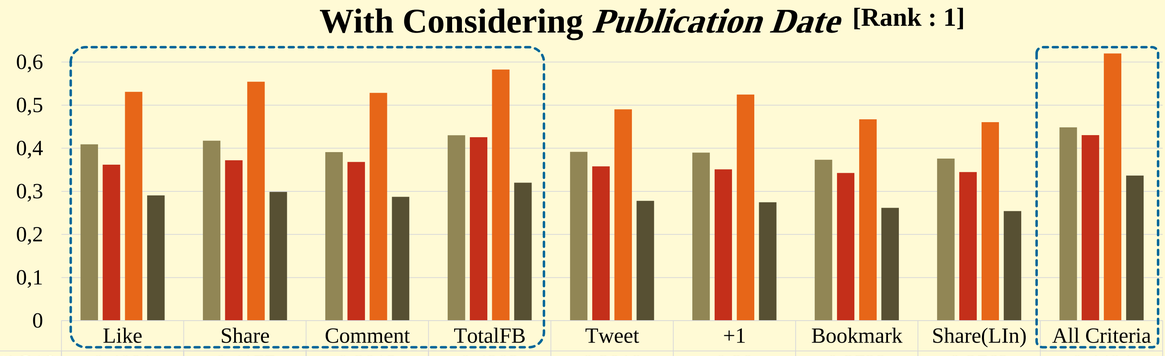
	Share	Comment
P@10	0,4148	0,3861
P@20	0,3681	0,3601
nDCG	0,5472	0,5207
MAP	0,297	0,2844

Baselines: Without Considering Time **[Rank : 3]**



	Like	Share	Comment	TotalFB	Tweet	+1	Bookmark	Share(LIn)	All Criteria
P@10	0,3938	0,4061	0,3857	0,4209	0,3879	0,3826	0,373	0,3739	0,4408
P@20	0,362	0,3649	0,3551	0,4102	0,3512	0,3468	0,3414	0,3432	0,4262
nDCG	0,513	0,5262	0,5121	0,5681	0,4769	0,5017	0,4621	0,4566	0,5974
MAP	0,2832	0,2905	0,2813	0,3125	0,2735	0,2704	0,26	0,2515	0,33

With Considering *Publication Date* **[Rank : 1]**



	Like	Share	Comment	TotalFB	Tweet	+1	Bookmark	Share(LIn)	All Criteria
P@10	0,4091	0,4177	0,3912	0,4302	0,3918	0,39	0,3732	0,3762	0,4484
P@20	0,362	0,3721	0,3683	0,4258	0,3579	0,3511	0,3427	0,3449	0,4305
nDCG	0,5308	0,5544	0,5285	0,5827	0,4903	0,5246	0,4671	0,4606	0,62
MAP	0,2907	0,2989	0,2874	0,32	0,2779	0,2748	0,2618	0,2542	0,3366