



**HAL**  
open science

# A Priori Relevance Based On Quality and Diversity Of Social Signals

Ismail Badache, Mohand Boughanem

► **To cite this version:**

Ismail Badache, Mohand Boughanem. A Priori Relevance Based On Quality and Diversity Of Social Signals. 38th Annual ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2015), Aug 2015, Santiago, Chile. hal-03154385

**HAL Id: hal-03154385**

**<https://hal.science/hal-03154385v1>**

Submitted on 28 Feb 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A Priori Relevance Based On Quality and Diversity of Social Signals

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

## 1. Introduction

### Context:

- Exploiting social signals to enhance a search.
- Do the quality and diversity of signals matter to capture relevant documents?

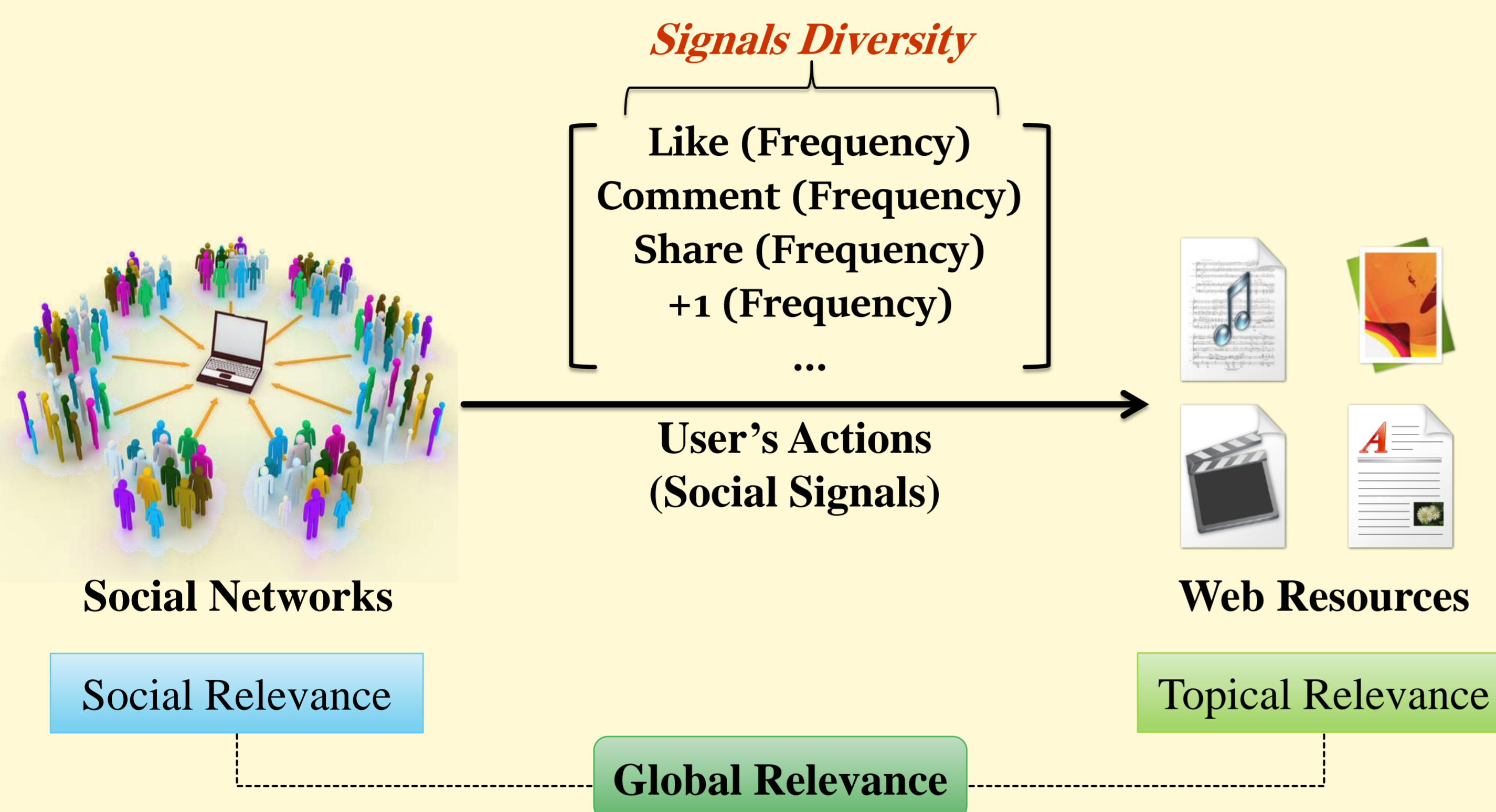


Figure 1. Global presentation of our approach

**Hypothesis 1:** Diversity of signals associated with a resource is a clue that may indicate an interest beyond a social network or a community, i.e., a resource dominated by a single signal should be disadvantaged versus a resource with an equitable distribution of the signals.

**Hypothesis 2:** Origin of social signals might impact the retrieval.

### Research Questions:

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

## 2. Social Signals Diversity

Using language model to estimate the relevance of document  $D$  to a query  $Q$ .

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

$P(D)$  is a document prior.  $w_i$  represents words of query  $Q$ .

- Signals are grouped according to their property  $x \in \{P: Popularity, R: Reputation\}$
- The priors are estimated by a counting of actions  $a_i$  associated with  $D$ .

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

Smoothing  $P(a_i^x)$  by collection  $C$  using Dirichlet:

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

Where  $P_x(D)$  represents the a priori probability of  $D$ .  $x \in \{P, R\}$  refers to the social property estimated from a set of specific actions.  $Count(a_i^x, D)$  represents number of occurrence of action  $a_i^x$  on resource  $D$ .  $a_i^x$  designs action  $a_i$  used to estimate  $x$  property.  $a_i^x$  is the total number of signals.

Estimating signals diversity in a resource using diversity clue of Shannon-Wiener.

$$Diversity_s(D) = - \sum_{i=1}^m P_x(a_i^x) \cdot \log(P_x(a_i^x)) \quad (4)$$

Where  $m$  represents the total number of signals.

The Shannon clue is often accompanied by Pielou evenness clue:

$$Diversity_s^{evenness}(D) = \frac{Diversity_s(D)}{MAX(Diversity_s(D))} = \frac{Diversity_s(D)}{\log(m)} \quad (5)$$

The general formula of  $P_x(D)$  becomes as follows:

$$P_x(D) = \left( \prod_{a_i^x \in A} P_x(a_i^x) \right) \cdot Diversity_s^{evenness}(D) \quad (6)$$

## 3. Experimental Evaluation

### Dataset:

- INEX IMDb Dataset.
- 30 INEX IMDb Topics and their relevance judgments.
- 7 social signals from 5 social networks.

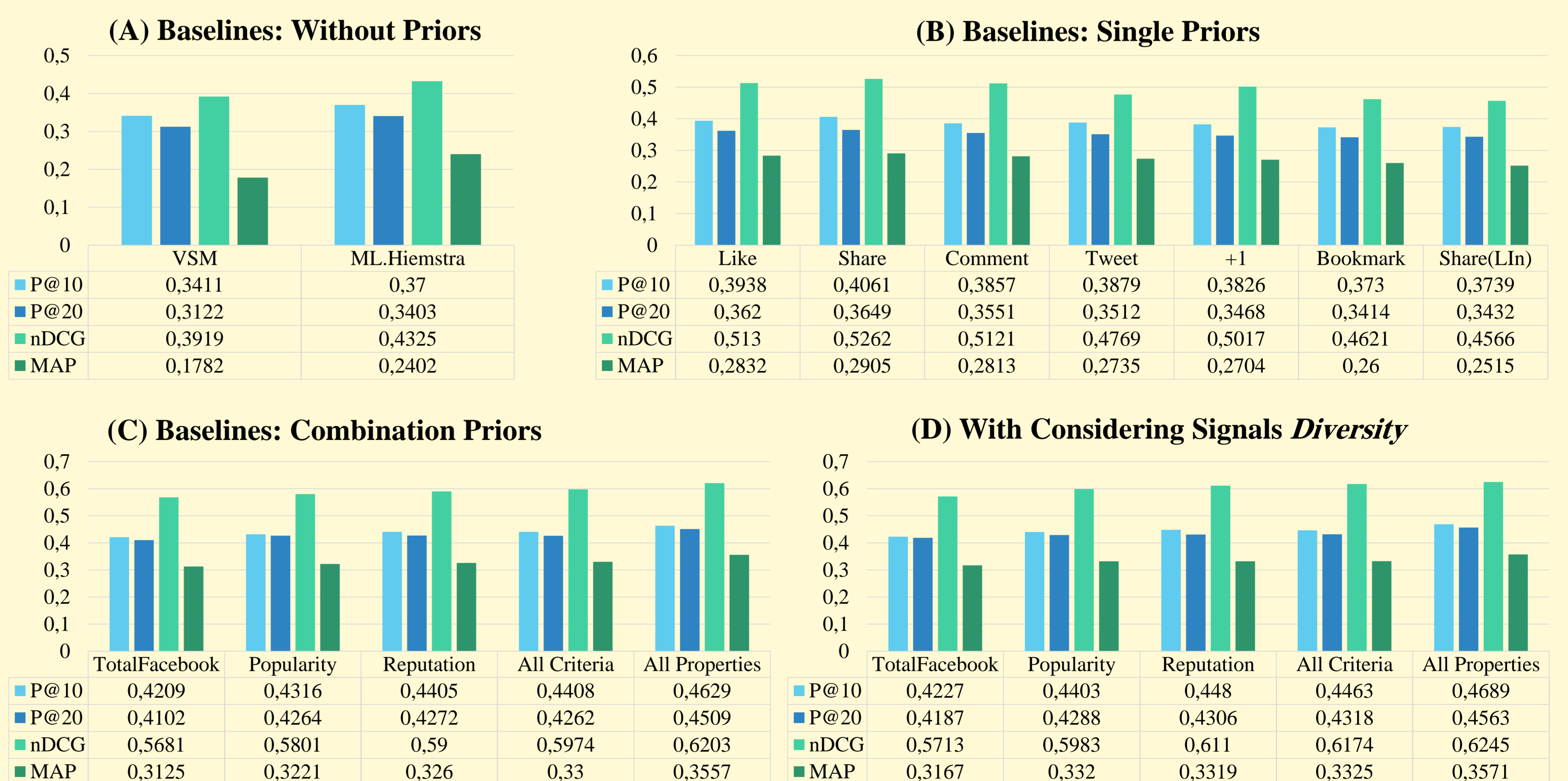
Table 1. Exploited social signals in quantification

Property	Social signal	Social 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 +1	Google+
	Number of Bookmark	Delicious

Table 2. Instance of document with social signals

Document id	Like	Share	Comment	+1
tt1730728	30	11	2	0
	Bookmark	Tweet	Share(LIn)	
	0	2	0	

### Results:



## 4. Quantitative and Qualitative Analysis

Figure 2. Signals % in the relevant documents

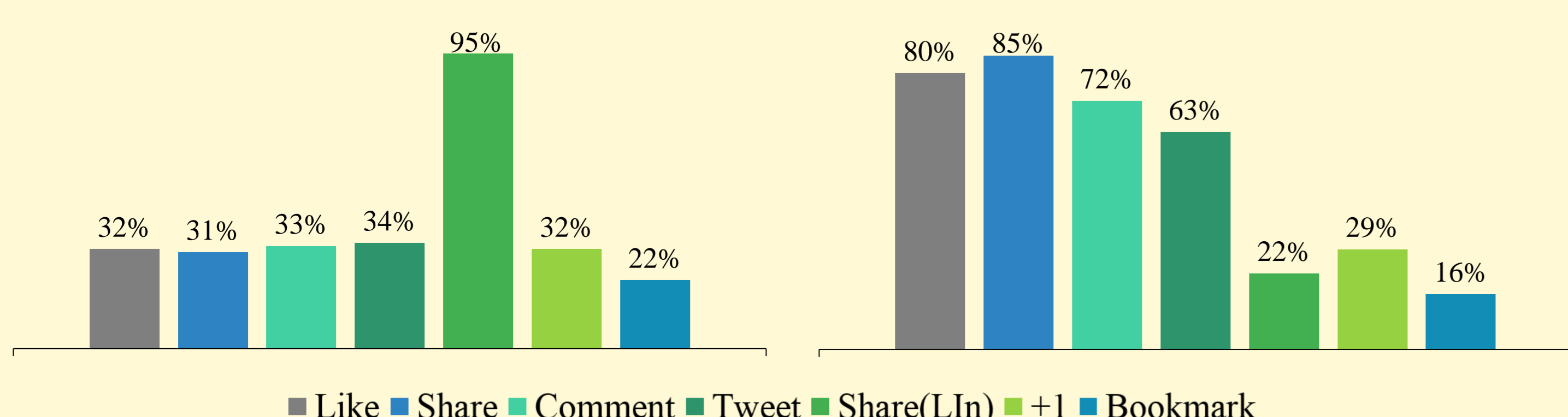


Figure 3. Relevant documents % containing signals

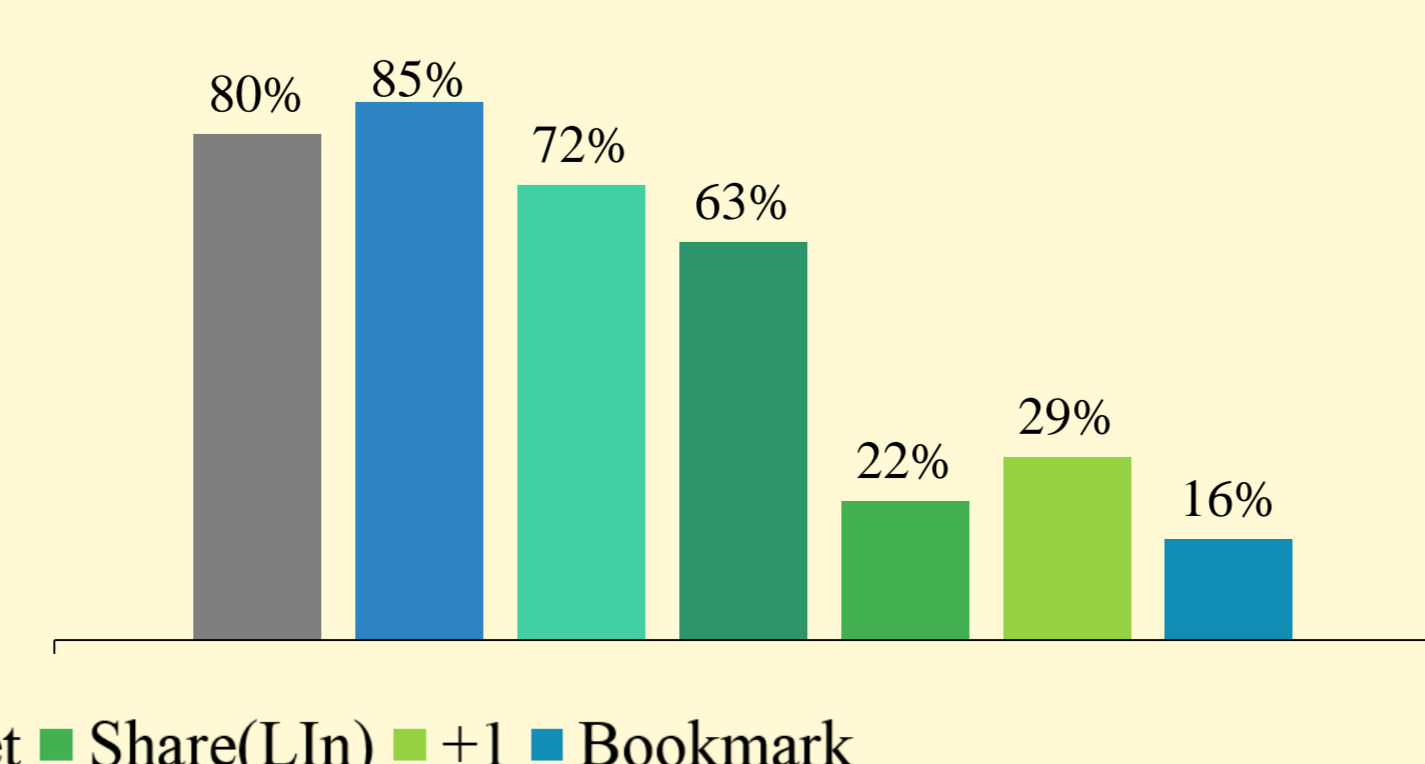


Table 3. Statistics on the distribution of the signals in the documents (relevant and irrelevant)

	Relevant documents containing signals			Relevant documents without signals		Irrelevant documents	
	Number of documents	Number of actions	Average	Number of documents	Number of actions	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: 2765			Total irrelevant: 27235				