International Conference on Control, Engineering & Information Technology (CEIT'13)
Proceedings Engineering & Technology - Vol.2, pp. 100-104, 2013

Social Relevance for re-Ranking documents of Search Engines Results

Amna Dridi
High Institute of Management
LISI Laboratory, Carthage University
Tunis, Tunisia
dridiamna@gmail.com

Hatem Haddad
Higher School of Science and Technology
LIPAH Laboratory, Tunis ElManar University
Tunis, Tunisia
haddad.hatem@gmail.com

***Abstract*—This paper presents an initial proposal for a formal framework that, by studying the social relevance, involved in information retrieval, can establish the re-ranking of search engines results and how to perform it. Social networks are used to find and to connect to other users but also to publish and to retrieve information. Traditionally, information retrieval uses the document content to fulfil users information needs. In the context of social networks, more information can be added to the document for example content annotation (tags). In this paper, we focus on social tags. We propose to use social tags to re-rank documents retrieved by a search engine. Experiments results on a documents collection of the Knowtex social network show that our approach can achieve better overall result compared with the traditional information retrieval approach.**

***Index Terms*—social information retrieval; social relevance; social network;**

I. Introduction

Web 2.0 technologies have led to the emergence of new social media: blogs, wikis, podcasts, file sharing platforms and social networks that we are concerned in this paper. Using social networks (Twitter, Facebook, linkedIn, or to Viadeo latest Foursquare, Gowalla), users moved from a passive state where they were information consumers to an active state where they are information producers. A new research domain is then created : social information retrieval (SIR). In this context, we propose a model for SIR which combines content relevance and social relevance to re-rank search results based on user personalized preferences.

This paper is organized as follows: Next, we summarize the related work in section II. Then, we describe our formal framework model of Social Information Retieval based on community detection in social networks in section III. We show the experiments and analyse the results in section IV. We conclude in section V.

II. Related work

Social Information Retrieval (SIR) is a new Information Retrieval (IR) domain that exploits social information produced by the social web (social networks, blogs, wikis ...) to customize the information search process and to retrieve relevant results corresponding to users information needs.

Social relevance is used in the state of the art in different ways. [1] combined the document content relevance with informations related to the document's author in order to calculate a new document relevance score. To evaluate the effectiveness of this approach, a series of experiments are conducted on a scientific document dataset that includes textual content and social data extracted from the academic social network CiteULike. Final results show that the proposed model improves the retrieval effectiveness and outperforms traditional and social information retrieval baselines. But the main limitation of this model is that social relevance is based on weighting social relationships which takes into account the authors positions in social network and their mutual collaborations. So that, we can find a relevant document which judged as non relevant because its author hasn't a central position in the social graph.

[3] proposed a model for social web search, called LAICOS, where the document index structure is structured into two parts. The first one is constructed using the document content and the second part is based on the document annotations. Experiments are conducted using documents indexed from the del.icio.us dataset and show the effectiveness of this model comparing with traditional Information Retrieval Systems (IRS). LAICOS ignores the information searcher which represents an important node in the social graph.

[11] used the ACT (Author- Conference - Topic) model that selects the five closest sub-topics to the query and then looks for the most influential authors. They have developed an influence maximization algorithm to find the sub network that closely connects the influential users. Two systems have been developed to evaluate this algorithm. The first system is deployed in Arnetminer.org and the other system is deployed in Tsinghua university centenary celebration system. Results confirm the effectiveness of this method by the largest Expand/Remove ratio comparing with the Random and the Path algorithm and also by the longest viewing time of a user on the returned social graph, but this algorithm is so time consuming. Another limitation of this model is the probability theory used to quantify relationships between authors and topics which is unable to express partial and total ignorance.

[8] designed Kodex, a system for detecting communities in a bipartite graph to automatically order Web search results by their relevance. Given a query, documents retrieved are modeled in a bipartite graph and then communities are extracted from this structure. The model disadvantage is that it's based on an initial partition choice and communities merging and results are affected by these two parameters.

[6] proposed a framework called SNDocRank that considers the documents contents and the relationship between the information seekers and the documents owners in a social networks. This approach combines the traditional tf-idf ranking measure and Multi-level Actor Similarity (MAS) algorithm that measures the structural similarity between the documents owners and the information seeker in a social network. This ranking method is implemented in simulated video social network extracted from YouTube. The results show that compared with traditional ranking methods the SNDRank algorithm returns more relevant documents. But the major limitation of this approach is that the effectiveness of the results depends on the social network, number of friends and local communities of the searcher.

[4]proposed an approach of query expansion based on the users profile. Informations from user's profile are added to user's queries considering the social proximity between the query and the user's prole. The proposed approach has been evaluated using a large dataset crawled from del.icio.us. Results show that this approach can perform better than the closest related work. The main limitations of this approach are its high level of subjectivity and the the problem of number of terms added to the query.

Our approach also is based on personnalizations principle where content and social relevance are combined to re-rank search results after the extraction of user community from a social network.

Various methods have been proposed to solve the problem of communities detection from social networks. [10] in his PhD thesis has cited the classical methods, separation methods, agglomerated methods, hierarchical clustering and he has focused on random walk method.

[2] has developed measures of centrality based on the shortest paths computation as Degree Centrality, Betweenness Centrality, Closeness Centrality and centrality measures based on steps of Bonacich power (eigen vector centrality). But the major limitation of these measures is that they detect communities based only on the structure and appearance of general network.

To solve this problem, [7] have used Jaccard coefficient to calculate the similarity between two users in Facebook based on social activities (link friendship, participation in groups...). In case of a null result, Jaccard coefficient has a the disadvantage of the similarity lack between two users whereas this is not true. To solve this problem, a popular parameter introduced by social science called ***Katz coefficient*** is used to calculate the similarity between two users taking into account all possible paths between two nodes.

We propose in our work to use Katz coefficient in order to detect communities in social network because of its effectiveness to take into account various types of links between two nodes in the social graph.

A. Katz Coefficient

Katz coefficient is a similarity index proposed in the field of social science and was recently reused in the context of collaborative recommandation and Kernel methods where they are known as Von Neuman Kernel. Katz proposed a method of calculating similarity taking into account not only the number of direct links between the elements, but also the number of indirect links [5].

Katz is the coefficient of the weighted sum of the number of direct paths between two nodes [9].

$$Katz := \sum_{l=1}^{N} \beta^l \mid {paths^l}_{i,j} \mid$$

with:

- l: length of the path
- β^l : the appropriate weight to the path l

III. Social Information Retrieval Model

Classical Information Retrieval Systems (IRS) are designed to retrieve relevant results corresponding to users information needs. Relevance score in this case is relative to document content so that relevance is called content relevance.

With the emergence of social networks, a new information type is occur with social tagging, user profiles and social activities. This information is called *social information*. Therefore, in this social context, the document can be socially evaluated according to **social relevance**.

In this article, we will detail our approach for SIR based on linear combinaition of content relevance and social relevance to re-rank search results.

We start by a step of classical IR where results are ordered according to content relevance and we reuse returned results in order to re-rank them according the social relevance.

A. Social Relevance

1) Social Information: In the context of web 2.0 and the emergence of blogs, wikis and social networks, user

became information producer. He annotates documents and web pages, he has different relationships with other users. He has a social prole. The information produced is so called **social information**.

Social information is, therefore, any information provided through the use of web 2.0. It's used to predict users interest and intentions. It's incorporated in the IR process to customize the search and gives the users the most appropriate answers to their information needs.

As the content relevance is a weighting relative to document content d_x, social relevance is a weighting relative to social activities related to the document d_x.

A document d_x belonging to SN_s (Social Network of similar users) can be a text document, an image, a video or multimedia document. It's defined by the following quadruplet (ct, l, s, c) where :

- ct (content): the content of document d_x
- l, s, c: social activities SA_i (l: like, s: share, c: comment) relative to the document d_x

In our case, the social relevance is the degree of popularity of the document d_x expressed by social activities related to d_x in the social network SN_s.

2) Social relevance computation: A document d_x has its social relevance in SN_s. Therefore, for the same query Q expressed by two different users U_x and U_y, returned results are ordered differently depending on the social context of each user.

Our aim is to focus on social relevance and to show how the integration of document social score ss_dx in the final document relevance score influences the re-ordering of SIR results. For [1], the social relevance is estimated using centrality measures: betweenness, closness, page rank, HITS authoriy score and HITS hub score.

Our idea for the social relevance computation is to find a social score for social activities which is the weighted sum of social activities weighted scores.

We consider the following social activities scores : $s1_{dx}$ (like's score), $s2_{dx}$ (share's score) and $s3_{dx}$ (comment's score), where:

- N(U_x) : is the total friends number U_x in the social network SNs.
- **like's score**: $s1_{dx}$=$\frac{p(l)}{1-p(l)}$ where p(l) = $\frac{l(Ux)}{N(Ux)}$ and l(U_x): the friends number U_x who clicked like for a document d_x in the social network SNs.
- **share's score**: $s2_{dx}$=$\frac{p(s)}{1-p(s)}$ where p(s) = $\frac{s(Ux)}{N(Ux)}$ and s(U_x): the friends number U_x who share a document d_x in the social network SNs.
- **comment's score**: $s3_{dx}$(c)=$\frac{p(c)}{1-p(c)}$ where p(c) = $\frac{c(Ux)}{N(Ux)}$ and c(U_x): the friends number U_x who comment a document d_x in the social network SNs.

We propose to combine the weighted scores of social activities $S_{dx}(SA_i)$ as follows :

$$\text{ss}_{dx} = \sum_{i=1}^{3} \alpha_i \; Si_{dx}$$

where: $S_{dx}(SA_i)$: the relative score of social activity SA_i $\sum_{i=1}^{3}\alpha_i$=1 ; α_i a weighted coefficient selected by the user U_x

B. Linear combination of content relevance and social relevance

In our model, we propose to combine social score ss_{dx} with content score that we called ssim (Q, d_x) (similarity score between the query Q and the document d_x) previously found by the IRS in a linear combination according a weighted coefficient λ chosen by the user U_x to find the final score S_{dx} of the document.

$$\text{S}_{dx} = \lambda \text{ ssim(Q, } d_x) + (1-\lambda)\; ss_{dx} \quad where 0 < \lambda ¡ 1$$

IV. Experimental Evaluation

In this section, we present our evaluation goals, a description of the dataset used, and experiments results.

A. Evaluation objectives

The experimental evaluation of our approach is undertaken through an hybrid evaluation combining content simulation and user study where we used real data issued from the scientific social network Knowtex[1]. Experiments are conducted to acheive the following objectives:

- Evaluate the impact of document social activities on the computation of its social relevance: the goal is to extract social activities related to each document (links) from Knowtex returned in classical search results in order to weight them and calculate relative social scores to have the final social score.
- Evaluate the effectiveness of the combination between content relevance and social relevance in order to reach a personalized re-ranking.
- Compare our approach of social ranking to classical approach by the assessment of satisfaction rate returned by real users.

[1] www.knowtex.com

TABLE I
STATISTICAL CHARACTERISTICS OF THE SOCIAL NETWORK KNOWTEX

Number of links	34017
weblists	1128
contacts	2492

B. Dataset

As there is no public available dataset for social search evaluation purpose, we exploited scientific papers (links) drived from social network, namely Knowtex, and we gathered data about social activities of each document. Before that, we asked each user to select his social graph of similar friends. Community detection is done manually by users with respect of Katz coefficient and the number of nodes in the social graph is limited to 10.

Social network properties: Knowtex is a community that explores sciences's culture, technology, design and innovation. It's developed by Umaps[2]. It organizes web resources (texts, videos, slides, etc.) collected by its members. Open to public in September 2009, Knowtex is a space for interconnected journalists, artists, mediators, designers, bloggers, researchers. Knowtex includes 34017 links, 1128 weblists and 2492 members with an average of 13 links per member (see Table I).

C. Effectiveness Evaluation of our model

The document collection consists of collecting the top 10 results retrieved from Knowtex for each testing query. Results are crawled and each one is represented by its complete content. We note that content relevance ssim(Q, d_x) is computed using the ranking function Okapi BM25. In our evaluation setting, these documents are used only for re-rank the search results using the social relevance. In order to evaluate the retrieval effectiveness, the relevance assessments for the testing queries were given through a user study. To do this, four Knowtex's members (Audrey Bardon, Civilisation2, Camilles, knowtex) were presented with the set of top 10 results retrieved from Knowtex. Each participant was considered as the user U_x who has formelated the query Q.

1) He was asked to choose the α_i weighting coefficients in order to weight social activities.

2) He was asked to choose the coefficient for weighting content relevance and social relevance.

3) He was asked to judge whether each document was correct or not according to the query and the preferred order.

[2]www.umaps.fr

TABLE II
CHOICE OF α_i TO WEIGHT SOCIAL ACTIVITIES

	α_1(weblist)	α_2(comment)	α_3(share)
Audrey	0.5	0.1	0.4
Civilisation2	0.4	0.3	0.3
CamilleS	0.3	0.2	0.5
knowtex	0.5	0.2	0.3
	0.425	0.2	0.375

4) We have asked each member to re-rank the top 10 results as he has wanted to have in order to estimate the satisfaction rate of returned results by classical model (content relevance), social model (only social relevance) and our model (linear combination between social relevance and content relevance).

We note that in Knowtex there are five social activities related to each document which are: evaluate by a member, add to weblist, comment, suggest a member and share. As we considered only three social activities SA_i in our approach, we took these three SA_i from Knowtex to respect our model: Add to weblist, comment and share.

The choice of α_i by the four members is represented by Table II. The taken into account to computing social scores α_i are the averages of the α_i choosen by the four users in order to generalize the evaluation.

After the computing of social scores, returned results are re-ranked by social relevance. We proposed an evaluation measure of user relevance that we called *satisfaction rate* which integrates the user judgement in re-ranking results. This meausure is defined as follows:

$$\text{Satisfaction Rate} = \frac{NumberOfEstimatedTrueAnswers(nbETA)}{TotalNumberOfAnswers(nbTA)}$$

We note that:

- nbTA: the total number of answers returned by the IRS
- RLO: Returned Link's Order by the IRS
- preferred order is the consultation order of returned results referred by PLO
- Estimated True Answer is a relevant answer in the preferred order and Number of Estimated True Answers (nbETA) is found according the following algorithm

Algorithm 1 Satisfaction Rate

Require: $nbTA \in N$, $nbETA \in N$, $RLO \in N$, $PLO \in N$
Ensure: $init(nbETA)$
FOR $i = 0$ to $nbTA$
if $[(RLO = PLO)||(|RLO - PLO| = 1)]$ **then**
$nbETA = nbETA + 1$
end if
END FOR
$RETURN nbETA$

TABLE III
CHOICE OF λ TO WEIGHT CONTENT RELEVANCE AND SOCIAL RELEVANCE

	λ	$\lambda - 1$
Audrey	0.2	0.8
Civilisation2	0.4	0.6
CamilleS	0.5	0.5
knowtex	0.6	0.4
	0.425	0.575

TABLE IV
RE-RANKING OF SEARCH RESULTS ACCORDING TO THE RELEVANCE SCORES

d_x	ssim(R, d_x)	r(d_x)	ss(d_x)	r(d_x)	S(d_x)	r(d_x)
d_1	0.501	1	0.099	4	0.1794	3
d_2	0.462	2	0.00	8	0.0924	8
d_3	**0.460**	**3**	0.00	9	**0.092**	**9**
d_4	0.453	4	0.045	7	0.1266	6
d_5	0.440	5	0.00	10	0.088	10
d_6	0.428	6	0.083	5	0.152	5
d_7	0.402	7	0.1494	2	0.19992	2
d_8	0.315	8	0.115	3	0.115	4
d_9	**0.300**	**9**	0.1814	1	0.20512	**1**
d_{10}	0.209	10	0.083	6	0.1082	7

Depending on satisfaction rates computed for each information retrieval model, the four users choose the weighting factor λ to weight content relevance and social relevance. Table III presents choosen λ by the four users.

We detailed in Table IV an exemple query expressed by the user Audrey Bardon which is scientific social network . We focused on d_3 and d_9 (document3 and document9) that had a remarkable rank changing which shows the big influence of the weighting factor λ on re-ranking results.

After each experiment, satisfaction rate is calculated 3 times: the first time is after the phase of traditional IR, the second time is after the using of social relevance only and the last time is when our approach is applied (the linear combination of two relevances is applied). Results show that the satisfaction rate of classical ranking results does not exceed 30% while satisfaction rate of social ranking results exceeds 60% and finally the satisfaction rate of combined classical and social ranking results is about 50%. Thus, compared to the classical model Okapi BM25, our model shows an important improvement that exceeds 20 %.

We can conclude that our approach of re-ranking search results based on social relevance in addition to content relevance improves the retrieval effectiveness compared to traditional baselines.

V. CONCLUSION

We proposed in this paper a social information retrieval model that combines content relevance and social relevance to re-ranking search results. Our model includes new social relevance definition which is based on social activities weighting that reflects social popularity of the document. Our experiments results on the Knowtex dataset reveals that satisfaction rate measure is able to better evaluate the user relevance and shows that our model can perform better that a traditional retrieval model.

In a futur work, more social activities related to the document will be integrated to social relavance computing. We plan also to conduct experiments on social document dataset that covers various research areas.

REFERENCES

[1] L. Ben Jabeur and L. Tamine and M. Boughanem, A social model for Literature Access: Towards a weighted social network of authors, CORIA, University Publication Center, pp. 403-404, 2010.

[2] C. Bothorel, Social network analysis and unpopular content recommendation, Review of New Information Technologies (RNIT), Vol. A.5, 2011.

[3] M. R. Bouadjenek and H. Hacid, LAICOS: A social web search engine, WW Panel CNRS, 2012.

[4] M. R. Bouadjenek and H. Hacid and M. Bouzeghoub and J. Daigremont, New Social approach for expansion query in web 2.0, CORIA, pp. 41-48, 2011.

[5] F. Fouss, and A. Pirotte, and J.M. Renders, and M. Saerens, Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation, IEEE Transactions on Knowledge and Data Engineering (TKDE), Vol.19, pp.2007, 2006.

[6] L. Gou and X. L. Zhang and H. H Chen and J. H. Kim and C.L. Giles, Social Network Document Ranking, JCDL '10 Proceedings of the 10th annual joint conference on Digital libraries, New York, NY, USA, pp. 313-322, 2010.

[7] P. De Meo And E. Ferrara And G. Fiumara, Finding Similar Users In Facebook, Social Networking And Community Behavior Modeling: Qualitative And Quantitative Measurement, IGI Global, pp. 304–323, 2011.

[8] E. Navarro and Y. Chudy and B. Gaume, Community detection in a bipartite graph and its application to the automatic classification of web search results (Kodex System), First day for models and network analysis: Mathematics and Computer Science Approaches: MARAMI, Toulouse, France, 2010.

[9] T.Y. Ouyang, Leveraging Temporal Features for Link Prediction in Communication Networks, Massachusetts Institute of Technology, DHS Summer Internship Report, 2007.

[10] P. Pons, community detection in large real-world graphs, PhD thesis, - Denis Diderot Paris-7 University, 2007.

[11] J. Tang and S. Wu and B. Gao and Y. Wan, Topic-Level Social Network Search, 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA, pp. 769-772, 2011.