Information Retrieval Framework based on Social Document Profile

Amna Dridi
supervised by: Mouna Kacimi

Faculty of Computer Science
Free University of Bozen-Bolzano I-39100, Italy
{Amna.Dridi, Mouna.Kacimi}@unibz.it

Abstract. Social networks provide rich information about user interests and activities representing a valuable source for search personalization. However, social information is typically large and dynamic making its exploitation to obtain relevant search results a very challenging task. This work presents a PhD project plan that investigates Social Information Retrieval. The goal is threefolds: (1) create confidence area for information search by community detection based on tags similarity (2) introduce a new notion of Social Document Profile based on user activities, and (3) propose a novel ranking model based on social relevance.

1 Introduction

1.1 Motivation

Social networks are becoming one of the predominant sources of information. Users of such networks publish documents that can take different forms, including text, image, audio, and video. Additionally, they can perform different types of actions around published documents. These actions can be classified as *descriptive* or *reactive*. Descriptive actions, mainly *tagging*, reflect the content of documents, while reactive actions such as *like*, *dislike*, *rate*, *favorite*, *share*, and *comment* reflect users' feedbacks regarding documents. This rich repository of users' actions triggered many research works to exploit social information for search personalization [3–5, 5, 5, 10, 12–14]. Most of the existing techniques consider descriptive actions (tagging) as the main indicator of users interests and thus use them for building users and documents profiles. However, relying only on tagging actions to provide relevant search results to users' needs is not sufficient. For example, a video tagged by {*Wolswagen*, *car*, *advert*} would be returned as a relevant result to the query *"car advert"* initiated by a user interested in *"Wolswagen"*. Knowing that the video features people speaking in fake Jamaican accents, some users would find it funny while some others would find it offensive. In this case, the video should be relevant only if it is liked by users having similar profiles to the query initiator. Consequently, the pool of users' reactions should be exploited to refine the search space and give a new definition for social document relevance. The contrast between descriptive actions

which are directly related to the content of documents and reactive actions that show users' personal preferences makes the exploitation of social information a challenging task.

1.2 Contribution and Paper structure

We propose to provide tailored answers to users' needs by exploiting social information in two different stages. First, we use descriptive actions to create, for each user, a confidence search area according to his profile. Second, we use both descriptive and reactive actions to define a social profile, per confidence area, for each document. The novel contribution by this paper has the following salient properties:

1. We model a social information retrieval framework as an undirected graph of social entities (User, Document, Tags and Clicks) where links represent entities relations generated in a social context, Tags represent descriptive actions, and Clicks represent reactive actions.
2. We exploit user profile as a tool for community detection based on Tags similarity. The goal is to establish a confidence search area for each user.
3. We propose a novel Social Document Profile based on a tripartite graph (Content, Tags, Clicks) that represents documents not only using their content but also their social profile given by Tags and Clicks.
4. We propose a novel scoring model that combines content relevance based on user profile and social relevance based on social document profile.

Our proposed approach goes beyond existing IR personalization techniques in several ways. First, it combines two areas: community detection in social networks and information retrieval. Second, unlike existing approaches, we define personalization approach based not only on user profile but also on document social profile. Third, none of the existing approaches takes into account clicks as social information defining document profile.

2 Related Work

Search personalization using social information has been investigated extensively. The first class of approaches limits social information to annotations or tags [3–5, 13]. For instance, Bouadjenek et. al., [4] use tags to build user profiles and then use those profiles for query expansion. The idea is to compute social proximity between each query and the profile of its initiator. Vellet et. al., [13] present two techniques that build user and document profiles. The first technique use a vector space model incorporating the concepts of tag inverse document frequency and tag inverse user frequency in folksonomy systems. By contrast, the second technique adapts the BM25 probabilistic model to user and document vectors. Similarly, Bouadjenek et. al., [3] propose a framework for social web search, called LAICOS, which construct document profiles based on their content and associated tags. Cai et. al., [5] examine the limitations of TF-IDF-based

models showing that using absolute term frequency favors active users against non-active users. Moreover, inverted document frequency is not necessary useful in indicating users' preferences on tags or how a document is relevant to tags. Thus, the authors use a Normalized Term Frequency (NTF) to indicate the preference degree of a user on a tag and thus construct user profile. Then, they perform search by matching users' profile and documents profile.

The second class of approaches exploits, in addition to tags, social relationships between users [1,6,9,10,12]. For instance, Carmel et. al., [6] re-rank search results based on friendship relationships among users. Schenkel et. al., [10] propose a top-k algorithm for social search and ranking with two dimensional expansions: semantic expansion that considers the relatedness of different tags and social expansion that considers the strength of relations among users. In the same context, Gou et al. [9] propose a framework called SNDocRank that considers documents content and the relationship between information seekers and documents owners by combining TF-IDF and Multi-level Actor Similarity (MAS) algorithm. Tang et. al., [12] selects the closest sub topics to the query and then looks for the most influential users. They have developed an influence maximization algorithm to find the sub network that closely connects influential users. Similarly, Ben Jabeur et. al., [1] define social scores based on users' relationships which depend on users' positions in the social network and their mutual collaborations.

All approaches described above focus on how to generate user profile using social information but none of them takes into account social document profile. In our work, we exploit user profile not at query time but to detect interest communities as confidence search areas. Moreover, we build a social document profile based on clicks which was not considered in related work. A work that went beyond using only tags and user relationships is by Wang et. al., [14] who define users' interests based on users' activities. However, the authors consider activities that are not related to documents but about social relationships such as subscription to groups. In our work, we use Clicks which are main indicators of documents social relevance.

Another research area related to our work is community detection where various methods have been proposed [2, 6, 7]. For instance, Bothorel et. al., [2] develop measures of centrality based on the shortest paths in social networks such as: Degree Centrality, Betweenness Centrality, and Closeness Centrality. De Meo et. al. [7] take a different approach than using network structure and propose Jaccard coefficient to calculate the similarity between users in Facebook based on social activities. In case of a null result, Jaccard coefficient has a 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. Carmel et. al. [6] consider similarity between two individuals according to common activity in the context of LC's [1] social software: co-usage of the same tag, co-tagging of the same docu-

[1] IBM Lotus Connections

ment, co-membership of the same community, or co-commenting on the same blog entry. The latter approach fits our needs but since we do not have access to the corresponding platform, we adopt Katz coefficient and use it as tool for community detection in social networks because of its effectiveness to take into account various types of links between nodes in the social graph.

3 Social Information Retrieval Framework

We define the Social Graph SG as a tuple SG = {U, D, T, C, A_1, A_2} where U = {u_1, ..., u_k}, D = {d_1, ..., d_l}, T = {t_1, ..., t_m} and C = {c_1, ..., c_e} are respectively the set of Users, Documents, Tags and Clicks. A_1 = {u_i, d_j, t_f} ∈ U × D × T is a set of annotations reflecting each user u_i tagging document d_j with tag t_f and A_2 = {u_i, d_j, c_r} ∈ U × D × C is a set of clicks reflecting each user u_i reacting to document d_j using click c_r (see Figure 1).

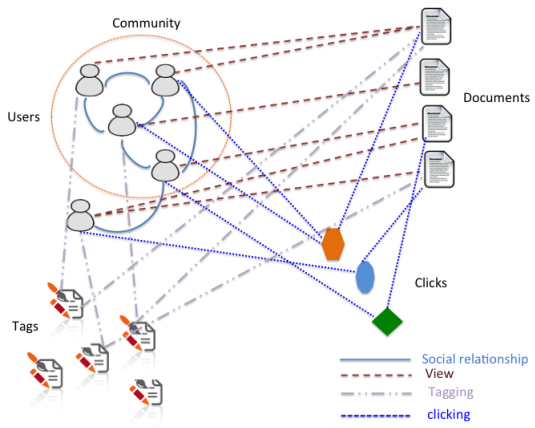

Fig. 1. Social Information Retrieval Graph

3.1 Overview

Our personalized search strategy consists in the following steps. First, we extract users' communities from social networks based on users' profiles. The profile of a user is defined by the set of tags he used to annotate documents. Thus, the community detection problem is reduced to computing tags similarity by using the subgraph G = (U, T) of the social graph SG. Second, upon receiving a search query $Q = \{q_1, ..., q_n\}$ from a user u, we proceed as follows:

1. We retrieve the topk relevant results to the query. Each result is associated with a content relevance score; the more relevant and important a result is with respect to the query, the higher its relevance score.

2. For each of the topk results, we compute its social score based on how popular it is in user u's community. This popularity is defined by related clicks (share, favourite, comment, etc).
3. The results are then re-ranked based on the combination of the content relevance score and the social relevance score of each result.

3.2 Social User Profile-based community detection

Social User Profile Our proposed model for social information retrieval is based on a central phase of community detection. Our aim is to detect community of interest to personalize IR processes. We propose to use the subgraph G = (U, T) of the social graph SG to detect similar users based on the tags they use. Note that, we take into account the time factor s since users' interest change over time. Therefore, the social user profile P_i of user u_i is defined by $P_i = \{t_1, \ldots, t_m\}^s$. To detect community between users it is then to compute tags similarity.

Community Detection We propose to adopt Katz coefficient for community detection. Katz coefficient is a similarity index proposed in the field of social science and was recently rediscovered in the context of collaborative recommendation 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 elements, but also the number of indirect links [8].

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

where l is the length of the path and β^l is the appropriate weight to path l.

3.3 Social Document Profile

Each document has a social profile defined by annotations (Tags) and Clicks in addition to its content. Therefore, a document D is defined by the threefold {Ct, T, C} where Ct, T and C respectively correspond to Content, Tag and Click. Therefore, a document is evaluated through two measures: *content relevance* and *social relevance*.

Content relevance. To compute the relevance of a document d_x to user query, we use BM25 (or Okapi) scoring function given by :

$$BM25(d_x, q_i) = IDF(q_i).\frac{f(q_i, d_x).(k_1 + 1)}{f(q_i, d_x) + k_1.(1 - b + b.\frac{|d_x|}{avgdl})}$$

where f(q_i, d_x) is the count of term q_i in document d_x, $|d_x|$ is the length of document d_x, avgdl is the average document length in the collection of documents,

$k_1 = 1.2$ and b = 0.75, IDF(q_i) is the inverse document frequency weight of the query term q_i which is computed as :

$$IDF(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

where N is the total number of documents in the collection, and n(q_i) is the number of documents containing n(q_i). Thus, the content relevance score of a document x is given by:

$$Rel(d_x, Q) = \sum_{i=1}^{n} BM25(d_x, q_i)$$

Social relevance. To compute *social relevance*, we use the tripartite graph (User, Document, Click) from the Social Graph SG. We consider the Clicks C = $\{c_1, \ldots, c_e\}$ to estimate the social popularity of a document in a given community. For the same query by two different users returned results are ordered differently depending on the social context of each user. Our idea for the social relevance computation is to to find a social score for clicks which is the weighted sum of clicks weighted scores. We consider the following click score of document d_x clicked by click c_i in the community of user u:

$$cs(d_x, c_i, u) = \frac{count(c_i, d_x, u)}{count(d_x, u)}$$

where: $count(c_i, d_x, u)$ is the number of users, in the community of user u, who used click c_i for document d_x, and $count(d_x, u)$ is the total number of users, in the community of user u, who clicked on document d_x. By combining the click scores, we obtain the social score of document d_x in the community of user u given by:

$$SS(d_x, u) = \sum_{i=1}^{e} \alpha_i cs(d_x, c_i, u)$$

where e is the number of clicks types (For example, in Facebook we have e=3 because we have 3 clicks types : like, share and comment) and $\sum_{i=1}^{n} \alpha_i = 1$ where α_i is a weighted coefficient selected by the query initiator.

3.4 Social Ranking Function

We use a linear combination of the content score $Rel(d_x, Q)$ and the social score $ss(d_x, u)$ to obtain the final score of a document d_x returned as a result for query Q initiated by user u:

$$S(d_x, u) = \lambda Rel(d_x, Q) + (1 - \lambda)SS(d_x, u)$$

where $0 \leq \lambda \leq 1$

4 Research Plan and Conclusion

As a short term objective, we plan to implement our personalized search approach and perform experiments on real-world data to evaluate its performance focusing on the following tasks: .

1. Compare our click-based personalization with tag-based personalization
2. Study closely the impact of the social document model on search results
3. Analyze how our technique performs depending on the level of activities in different communities.

4.1 Experimental Data

We will test our personalized search approach using data crawled from YouTube [2] which has the main characteristics needed for our solution. This dataset have been crawled during the period between October, 15th, 2012 and December, 25th, 2012. It contains 890682 videos, 282074 users and 1014190 information about social clicks (comment, favourite and rated).

Table 1. Statistical characteristics of YouTube dataset

Users	282074
Videos	890682
Clicks	1014190

4.2 Research Plan.

Our long term objectives consist in the following:

1. Investigate new techniques for community detection that go beyond tag similarity by involving users' reactions to published documents in social networks. We believe that building confidence search areas based on what users think about documents is a promising direction towards satisfying user's needs.
2. Extend the notion of document social relevance by considering not only positive feedbacks but also negative ones. The idea is to boost documents social scores if they receive positive feedbacks and penalize them otherwise. This task involve mainly mining users' comments to understand their interests and derive their judgment about published documents.
3. Develop an efficient and scalable ranking algorithm that can handle the fast growth of communities and the very high rate of content production together with tagging and clicking actions.

[2] www.youtube.com

4. Validate our proposed techniques using real datasets from social networks. We aim at investigating networks with different properties such as, Facebook, Twitter, and Delicious to understand the behavior of our approach is different environments.

Acknowledgements. This research was supported by the RARE project at KRDB research centre for knowledge and data at Free University of Bozen-Bolzano.

References

1. Ben Jabeur, L., Tamine, L., Boughanem, M.: A social model for Literature Access: Towards a weighted social network of authors. In Proceeding of RIAO 2010, pp. 32-39
2. Bothorel, C.: Social network analysis and unpopular content recommendation. Review of New Information Technologies (RNIT) 2011, Vol. A.5
3. Bouadjenek, M.R., Hacid, H., Bouzeghoub, M.: LAICOS: An open source platform for personalised social web search. In Proceeding of KDD 2013, pp. 1446-1449
4. Bouadjenek, M.R., Hacid, H., Bouzeghoub, M., Daigremont, J.: Personalized social query expansion using social bookmarking systems. In Proceeding of SIGIR 2011, pp. 1113-1114
5. Cai, Y., Li, Q.: Personalized Search by Tag-based User Profile and Resource Profile in Collaborative Tagging Systems. In Proceedings of CIKM 2010, pp. 969–978
6. Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'el, N., Ronen, I., Uziel, E., Yogev, S., Chernov, S.: Personalized Social Search Based on the User's Social Network. Proceedings of CIKM 2009, pp. 1227-1236
7. De Meo, P., Ferrara, E., Fiumara, G.: Finding Similar Users in Facebook, Social Neworking and Community Behavior Modeling: Qualitative and Quantitative Measurement. IGI Global 2011, pp. 304-323
8. Fouss, F., Pirotte, A., Renders, J.M., Saerens, M.: Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. In Proceeding of TKDE 2006, Vol.19, pp. 2007
9. Gou, L., Zhang, X.L., Chen, H.H., Kim, J.H., Giles, C.L.: Social Network Document Ranking. Proceedings of JDCL 2010, pp. 313-322
10. Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J.X., Weikum, G.: Efficient Top-k Querying over Social-tagging Networks. In Proceedings of SIGIR 2008, pp. 523-530
11. Ronen, I., Shahar, E., Ur, S., Uziel, E., Yogev, S., Zwerdling, N., Carmel, D., Guy, I., Har'El, N., Ofek-Koifman, Sh.: Social Networks and Discovery in the Entreprise (SaND). In Proceedings of SIGIR 2009, pp. 836-836
12. Tang, J., Wu, S., Gao, B., Wan, Y.: Topic-level Social Network Search. In Proceedings of KDD 2011, pp. 769-772
13. Vallet, D., Cantador, I., Joemon, M.J.: Personalizing Web Search with Folksonomy-based User and Document Profiles. In Proceedings of ECIR 2010, pp. 420-431
14. Wang, Q., Jin, H.: Exploring Online Social Activities for Adaptive Search Personalization. In Proceedings of CIKM 2010, pp. 999-1008