Query Expansion Based On Both Long-Term And Short-Term Social Relation In XML Social Book Search

Li Xinfang*
Tangyuan county power supply company
Heilongjiang electric power company limited
Tangyuan,China
e-mail: lxyyjh@126.com

Chen Qiang
Tangyuan county rural electrification bureau
Tangyuan,China
e-mail: cehn1234@163.com

Abstract—Query expansion technology is usually used to make abundant query to search information. Traditional query expansion based on WordNet, user search log or user feedback couldn't get the user's real intent, sometimes the latter could expand overly. In this paper, we propose a novel query expansion method based both on the user's long-term relation and short-term relation in XML social book search. We first propose recommendation degree of a tagged word based on a user's long-term social relation and short-term relation respectively, and the combined recommendation degree of a tagged word, then based on these definitions, we propose a new query expansion algorithm. The experimental results on real-world data show that our method has greatly increased the retrieval precise and users' satisfaction .

Keywords- query expansion; long-term social relation; short-term social relation; book search; recommendation degree

I. INTRODUCTION

More and more people search book information through internet now, this is usually realized by using a search engine. In general, the common user input short queries(for example, one word) to retrieve, but short queries are difficult to express the user's query intentions completely and the search engine would return many unrelated results. To improve the searching effect, query expansion is used to characterize the user's query intent. Traditional query expansion methods may be divided into two main kinds, one is based on a knowledge resource such as WordNet [1-2], the other is based on user search log [3] or user feedback [4]. Query expansion based on a knowledge resource focuses on finding the words that are meaningfully similar with or are hypernyms / hyponyms of the original query words; this could sometimes improve the recall ratio but yet couldn't get the user's real intent pages especially when the query words are not clear. Query expansion based on user search log or user feedback uses statistic feature between query words and other words appeared in all users' search history, but sometimes it could expand overly [5].

Nowadays, Social network supply an important role, through the use of micro-blogs, such as Twitter, people could release and exchange their opinions[6]. [7] found that social interactions play an important role throughout the search process.The advent of online social book community platforms also influences user's searching way. On the social book platforms such as Amazon or LibraryThing, there are many book tags and social tags. These tags supply extra information to use to search books. For example, the book tags could reflect the book's content; the social tags such as a user's friends or interested persons may influence the user‘s query intent to some extent. We think of a user's friends as their fixed social relation and the current interested persons as their temporary social relation, discuss the different effects for each role that influence the user's search intent and use these two kinds of user's social relation with their book tags to get the user's query intent so as to improve the searching effect.

In view of the above analysis, we propose a novel query expansion method based on the user's fixed relation and temporary relation in XML social book search.

II. RELATED RESEARCH

One main query expansion method is based on a knowledge resource. In reference [1-2], WordNet was employed to find the synonyms for a query word. [8] proposed query expansion based on the morphological thesaurus established. [9] expanded query based on Wikipedia and [10] based on formal domain ontology. These methods focuses on finding the words that are meaningfully similar with or are hypernyms / hyponyms of the original query words, however, the expansion words are not sure to reflect the user's real search intent. The other main query expansion method is based on user feedback or user search log. Pseudo-relevance feedback (PRF) exploiting the retrieval result has been the effective query expansion method[4]. [3] chosed query logs as data sets and then extracted the word most relevant with the new query as an extension word, [11] extracted the words relevant to the query word directly from the high-frequency-click results. This kind of query expansion sometimes could expand overly.

There is growing interest in social book search now. [12] re-ranked the searching results by using book tags, [13] ranked the searching results based on users' evaluation score and the author information. However, these researches mainly pay attention to the result ranking.

Social tags may also be used in query expansion. There are some researches on query expansion using social tags in other application fields. [14] proposed query expansion based on social tags for personalized web search, [15] extend expression of queries using social tags, [16] discovered useful matches between multilingual tags and applied such matches to expand user queries to retrieve additional resources tagged by the other languages. Among these researches, the social relation with the user here is too wide, the expanded query sometimes may deviate user's present query intent.

Different from above researches, our method limits the social relation on user's long-term relation--friends and short-term relation---interested persons, whose book tags may reflect the user's intention more effectively.

III. QUERY EXPANSION BASED ON BOTH LONG-TERM SOCIAL RELATION AND SHORT-TERM SOCIAL RELATION

A. Related definition

*1) User's social relation:*There are many kinds of social relation for a user, we mainly focus on user's friends and current interested persons. We think that a user's friends are those persons who have common or similar views and similar language custom with the user when they make tagging. And more, they may influence the user's selection to some extent. The user is familiar with his/her friends at a long time, so we think of this social relation as long-term relation. While a user's interested persons are those persons whose reading interests are similar with the user's recent reading interests, we call this kind of relation as short-term relation. Based on these social relation's book tags, a user's query intent could be inferred effectively.

2) User's book tags: The tags of books annotated by users are based on individual understanding to books, which are the summary of content and topics. User and his friends or interested persons all may annotate books he saw. The following list is the book tags for user j*:*

Book1(Sea poacher, anime)
Book2(anime, history)

3) Recommendation degree for a tagged word based on long-term relation: Considering different user's different book tags, we define the recommendation degree of a tagged word or phrase t for original query word q based on long-term relation as the following:

$$RD_F(q,t)=\frac{1}{N}\sum_{U_{F,q}}\frac{C_{q,t}}{C_q+0.001} \tag{1}$$

In (1), $C_{q,t}$ represents the times that q and t co-occur in the book tag set of a friend; C_q represents the times of q occurs in the same book tag set; to avoid the denominator to be zero, we add 0.001 to it. $U_{F,q}$ means those friends of a query user that all have used q as tagged word in their book tag set, and N is the total number of $U_{F,q}$, $RD_F(q,t)$ is the average recommendation degree for t of all friends in $U_{F,q}$,

4) Recommendation degree for a tagged word based on short-term relation: We define the recommendation degree of a tagged word or phrase t for original query word q based on short-term relation as the following:

$$RD_T(q,t)=\frac{1}{N}\sum_{U_{T,q}}\frac{C_{q,t}}{C_q+0.001} \tag{2}$$

In (2), $C_{q,t}$ represents the times that query word *q* and tag word *t* co-occur in the book tag set of an interested person; C_q represents the times of the original query word *q* occurs in the same book tag set; $U_{T,q}$ means a query user's interested person set that all have used *q* as tagged word in their book tag set, and *N* is the total number of $U_{T,q}$. $RD_T(q,t)$ is the average recommendation degree for *t* of all interested persons in $U_{T,q}$

5) combined recommendation degree for a tagged word: We define the combined recommendation degree of a tagged word or phrase t for query word q as the following:

$$RD_C(q,t)=r_1\times RD_F(q,t)+r_2\times RD_T(q,t) \tag{3}$$

r_1 and r_2 are the weight value satisfying $r_1+r_2=1$.

B. Query expansion algorithm based on both kinds of social relation

Input: account number of a user Ui and his original query q

Output: new query Q after expansion

The algorithm:

```
(1) Get the long-term relations set(friends) F1 and the short-term relations set(interested objects)F2 from the profile of $U_i$;
(2) For each object j in F1
    Find his/hers history book tags including query q and group them into a set denoted as $BTF_j$={book1(q,$t_2$,…,$t_m$), book2(q,$t_2$,…,$t_n$),……};
(3) Let BTF={$BTF_1$, $BTF_2$, …$BTF_j$, …};
    For each tag t in BTF
        For each $BTF_j$ in BTF
            Let $C_q$=the book count of $BTF_j$
            Let $C_{q,t}$ = the occurrence times of t;
        Calculate the $RD_F$(q,t) according to (1);
(4) For each object j in F2
    Find his/hers book tags including query q and group them into a set denoted as $BTT_j$={book1(q,$t_2$,…,$t_m$), book2(q,$t_2$,…,$t_n$),……};
(5)Let BTT={$BTT_1$, $BTT_2$, …$BTT_j$, …};
    For each tag t in BTT
        For each $BTT_j$ in BTT
            Let $C_q$=the book count of $BTT_j$
            Let $C_{q,t}$ = the occurrence times of t;
        Calculate the $RD_T$(q,t) according to (2).
(6) For each tag t in BTT or BTF
    If t doesn't occur in BTT then its $RD_T$(q,t) equals to zero, or, if t doesn't occur in BTF then its $RD_F$(q,t) equals to zero.
    Calculate the RD(q,t) according to (3).
(7) Select the first few tags with high RD(q,t) as expanding words and add them to the original query to rebuilt new query Q
```

IV. Experiments and results

A. Evaluation metric

We use nDCG (the normalized discounted cumulative gain) to evaluate the effectiveness of our query expansion method. Its definition is as the following [17]:

$$nDCG[k] = \frac{DCG[k]}{DCG'[k]} \qquad (4)$$

$$DCG[k] = \sum_{i=1}^{k} \frac{G[i]}{\log_2(1+i)} \qquad (5)$$

$G[i]$ is the correlated value with the query topic for the *i*th result returned by the search engine. DCG[k] is the discounted cumulative gain at rank k, DCG'[k] is the ideal discounted cumulative gain at rank k, here, we let k equals to 10, that is, we select the top ten searching results to calculate. The higher value of the nDCG@10 represents the higher user satisfaction and the more effective the words expanded are.

B. Experiment Data sets

The data sets we used are provided by the organizer of INEX 2012 Book Track. The book records consist of a collection of 2.8 million records from Amazon Books and LibraryThing.com. Each book record is an XML file with curated metadata and social metadata. The user profiles supply user's friends, interested persons, date and tagged books and other information; the query topics are taken from the LibraryThing discussion forums. We select one key word from each topic as the original query word.

C. Experiment process description

According to the query expansion algorithm proposed, we expand the original query word first. And then, we input the new expanded query into the retrieval system to check the results. We use nDCG@10 as our evaluation metric.

To see the different effect of the two different social relations, we modify the value of r1 and r2 in (3), let r2 change from 0 to 1, then the value of r1 will change from 1 to 0. We find that when r2=0.7, r1=0.3, we get the biggest nDCG@10. This implies that the impact of the user's short-term relation is a little stronger since the reading interest of a user's short-term relation may be same as the user's current reading interest. But when r2 is greater than 0.7, the nDCG@10 decreases. This implies that the impact of the user's long-term relation is also necessary; the two kinds of a user's social relation both infect the user's query intent.

To verify the effectiveness of our query expansion method, the original query, query expansion based on user's query log (here, user's query log means user's tagged history of books) and query expansion based on WordNet are used for comparison. After dropping those incomplete data, we select ten users' query topics from the LibraryThing discussion forums to check the effect of the four query expansion methods. Since there are many books with fewer than three tags in users' tagged book data, we expanded one and two words respectively to the original query. In our expansion method, the weight parameters we used are: r2=0.7, r1=0.3. The selected query topic words are listed in the following table.

TABLE I. ORIGINAL QUERY TOPICS

Topic ID	topic words
1	iliad
2	fantasy
3	history
4	crafts
5	japan
6	america
7	bioethical
8	thriller
9	Charl Dawin
10	Time travel

D. Experimental results and Discussion

We compare the retrieval effect of four different query expansion methods. Here, we use the normalized discounted cumulative gain at rank 10 (nDCG@10) [17] to evaluate the retrieval effects. The higher value of the nDCG@10 represents the higher user satisfaction and the more effective the query expansion is. The experiment results are shown in Table 2. According to Table 2, the nDCG@10s obtained by our method are all greater than that by original query, query expansion based on user's query log and by WordNet expansion. For method based on user's query log, the nDCG@10s for many topics are the same as the nDCG@10s for the original query. This is because that when a user's current query interesting isn't the same as before or there are few query logs for the user, we could find no words to expand the new query according to the user's query log. Even more, when a user's current query interesting isn't the same as before, the expanded query may reduce the searching accuracy, as for topic 5. For method based on WordNet, when the query words are person name such as topic 9, since we couldn't find any expanded words from WordNet, so the nDCG@10s are the same as the original query. The same case is the topic 10, the nDCG@10s are also the same as the original query

V. Conclusions

When searching book information with search engine through Internet, common users could not describe query requirements well for the limitation of professional knowledge and they usually use short query which would get poor searching results; traditional query expansion technology such as expansion based on user's query log or expansion based on WordNet can't reflect different user's query intention well. In this paper, we propose a new query expansion method based on both a user's long-term social relation and short-term social relation for social book search. The experimental results on real-world data show that our method has greatly increased the retrieval precise and users' satisfaction. Therefore, our method could get better results by making use of the wisdom of the user's social relationship. The experiment results also indicates that our method reflects the user's real-time interesting more effectively and may recommend different query words for different users through the analysis of the interest of the users' social relation. Next we would check our method in other searching field.

TABLE II. THE NDCG@10S OF SEARCHING RESULTS BY USING THREE EXPANDED QUERY METHODS

Topic ID	**Original query**	**Expansion based on user's query log**		**Expansion based on WordNet**		**Our expansion method**	
		Expanding number		**Expanding number**		**Expanding number**	
		1	**2**	**1**	**2**	**1**	**2**
1	0.6010	0.6010	0.6010	0.6010	0.5769	**0.8148**	**0.6859**
2	0.2541	0.5075	0.5477	0.3421	0.3134	**0.8014**	**0.8148**
3	0.1335	0.1335	0.1335	0.2365	0.2124	**0.5877**	**0.5877**
4	0.5476	0.5877	0.5877	0.4547	0.5138	**0.8014**	**0.6411**
5	0.4942	0.1736	0.3740	0.3876	0.5643	**0.6144**	**0.7128**
6	0.4744	0.4744	0.4744	0.6001	0.5877	**0.7079**	**0.6812**
7	0.4277	0.4277	0.4277	0.6376	0.5476	**0.6875**	**0.6810**
8	0.1274	0.1274	0.1274	0.3404	0.1335	**0.4408**	**0.4320**
9	0.3312	0.3312	0.3312	0.3312	0.3312	**0.4468**	**0.4468**
10	0.1253	0.2404	0.4408	0.1253	0.1253	**0.8484**	**0.7614**
average	**0.35164**	**0.3738**	**0.4045**	**0.4057**	**0.3906**	**0.6751**	**0.6445**

REFERENCES

[1] F.B.D.Paskalis and M.L.Khodra, "Word sense disambiguation in information retrieval using query expansion," *2011 International Conference on Electrical Engineering and Informatics*, Indonesia, 2011.

[2] M.Dragoni,C. DA Costa Pereia and A.G.B. Tettamanzi, "A conceptual representation of documents and queries for information retrieval systems by using light ontologies," *Expert System Application,* Vol. 39, No.12, 2012, pp. 10376–10388.

[3] Z.Yin, M. Shokouhi, and N. Craswell, "Query expansion using external evidence," *Proceedings of the 31th European Conference on Information Retrieva1,* Toulouse, France, 2009.

[4] J.H.Wang and M.H. Shih, "Relevant Term Suggestion Based on Pseudo Relevance Feedback from Web Contexts," Lecture Notes in Computer Science, Springer, Vol. 7634, 2012, pp.317-320.

[5] C.X.Wan and Y.Lu, "Structural query expansion based on weighted query term for XML documents," *Journal of Software,* Vol.19,No.10,2008, pp. 2611−2619.

[6] T.G.Arturo, S.M.Miquel and M. P.Josep, "Discovering social structures of local influence by using tweet Stimuli," *International Journal Computer Mathematics,* Vol. 12, 2013, pp. 1-13.

[7] B.M. Evans and E.H. Chi, "An elaborated model of social search," *Information Processing & Management,* Vol.46, No.6, 2010, pp.656-678.

[8] L. Araujo, H. Zaragoza, J. R. Perez-Agüera, and J. Perez-Iglesias, "Structure of morphologically expanded queries: A genetic algorithm approach," *Data & Knowledge Engineering,* Vol.69,No.3,2010, pp.279–289.

[9] T.T. He and X.L. Dai, "Pseudo-relevance feedback query based on Wikipedia," *2012 IEEE International Conference on Granular Computing*, Hangzhou, China , 2012.

[10] N.Alejandra Seguran,S.Salvador , G.B.Elena and M.E. Prieto, "An empirical analysis of ontology-based query expansion for learning resource searches using merlot and the gene ontology," *Knowledge-Based System*, Vol.24,2011 ,pp.119–133.

[11] S.Riezler, A.Vassermana, I.Tsochantaridis, V. Mittal, and Y. Liu, "Statistical machine translation for query expansion in answer retrieval," *Proceedings of the 45th Annual Meting of the Asscoiation For Computational Linguistics*, Prague, 2007.

[12] R.Deveaud,E.Sanjuan,and P.Bellot, "Social recommendation and external resources for book search," *INEX 2011: Social Book Search Track*,.Rome, Italy,2012.

[13] T.Bogers, K.W.Christensen, and B.Larsen, "*Focused Retrieval of Content and Structure*," Springer Berlin Heidelberg, 2012.

[14] C.Biancalana and A.Micarelli, "Social tagging in query expansion: A new way for personalized web search," *International Conference on Computational Science and Engineering,* Vancouver, 2009.

[15] C.Zhao, Z. Zhang, Xie X, and T.Liang, "A new keywords method to improve web search," *12th IEEE International Conference on High Performance Computing and Communications*, Melbourne, 2010.

[16] J.J. Jung, "Cross-lingual query expansion in multilingual folksonomies: A case study on Flickr," *Knowledge-Based System*, Vol.42,2013, pp. 60-67.

[17] S. Buttcher, L.A.C.Charles, and V.C.Gordon, "*Information retrieval:Implementing and evaluating search engines*," The MIT Press, 2010.