Building User Interest Profiles from Wikipedia Clusters

Jinming Min
Centre for Next Generation Localisation
School of Computing, Dublin City University
Dublin 9, Ireland
jmin@computing.dcu.ie

Gareth J. F. Jones
Centre for Next Generation Localisation
School of Computing, Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

ABSTRACT

Users of search systems are often reluctant to explicitly build profiles to indicate their search interests. Thus automatically building user profiles is an important research area for personalized search. One difficult component of doing this is accessing a knowledge system which provides broad coverage of user search interests. In this work, we describe a method to build category id based user profiles from a user's historical search data. Our approach makes significant use of Wikipedia as an external knowledge resource.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search and Retrieval

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

user profile, clustering, Wikipedia

1. INTRODUCTION

Personalization is a significant research area in information retrieval. Capturing information about user search interests is an important component in personalization. However, users are generally reluctant to explicitly create profiles of their search interests. Much research has been carried out to examining methods to automatically create user profiles to model the user search interests [5, 3, 7, 1]. One component of this is to identify a general knowledge system to provide good coverage of user search topics which can then be incorporated in models of user interests. In this work, we describe a method to build user profiles from a user's historical search data. In doing this, our methods make use of Wikipedia as an external knowledge resource.

The remainder of this paper is structured as follows: Section 2 introduces our system framework; Section 3 introduces the method to transform the Wikipedia collection into a general knowledge system with category information; Section 4 presents the algorithm to build user profiles from their historical search data; Section 5 describes our experimental setup and results, and we conclude this work in Section 6.

SIGIR 2011 Workshop on Enriching Information Retrieval (ENIR 2011), July 28, 2011, Beijing, China.

2. SYSTEM FRAMEWORK

Our research aims to build user profiles from the user's historical search data using Wikipedia. The system framework is presented in Figure 1. The system starts from the user's historical search data: historical queries and click-through documents. It is obvious that these data indicate the user's search interests. The user click-through documents usually contain many noisy terms. We utilize document reduction (DR) to select important terms from these documents [2]. The historical query terms and important terms from click-through documents are combined as user interests terms to form a query. The query is sent into a knowledge system to conduct retrieval. The knowledge system contains N categories clustered by topics, each category contains various numbers of documents. Using a text retrieval algorithm, a ranked list is produced. The top of each ranked document list is associated with a category id from the knowledge system. Thus these category ids can be formed into a vector which is our final user profile based on category information.

3. TRANFORM WIKIPEDIA INTO KNOWLEDGE SYSTEM

To model the user search interests, a well organized knowledge system with category information is needed. Wikipedia already has a large amount of category information for every document. In one category system, Wikipedia documents are divided into twelve broad categories: reference, culture, geography, health, history, mathematics, nature, people, philosophy, religion, society, technology. However a user's search interests will be more specific than these broad categories. Thus these high level categories are not sufficient to model a user's search interests. We propose to use a clustering algorithm to group the Wikipedia documents into categories. To do so, we use a k-means clustering algorithm to cluster the Wikipedia documents [6]. The k-means clustering procedure is as follows:

1. The first document processed is placed in the first cluster.
2. Each document in the collection is compared to each existing cluster and assigned to the highest scoring cluster that exceeds the specified threshold score.
3. If no cluster score exceeds the threshold, the document is placed in a new cluster.
4. Repeat steps 2 and 3 until all documents have been assigned to clusters.

To compute the similarity of documents in step 2 we use a standard cosine similarity shown in Equation 1. In Equation 1, the documents A and B are described by vectors including the term frequency of n terms. n is the total number of individual terms in the collection vocabulary.

$$similarity = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

After clustering, each Wikipedia document belongs to a category. Each category can be automatically assigned a *category id*, thus each Wikipedia document is associated with a *category id*.

4. BUILDING USER PROFILES

Given a user, his historical search data includes a historical query set Q(q_1,q_2,...,q_m) and click-through documents set D(d_1,d_2,...,d_n). For each click-through document d_j ($1 \leq j \leq n$), document reduction is carried out which calculates the Okapi BM25 weights to rank the importance of the terms in d_j. We use the top terms in d_j as key terms to indicate the search interests of the user. All the terms in q_i ($1 \leq i \leq m$) and key terms in d_j are combined into a union U. U may contain repeated terms since q_i and d_j may contain the same term twice or more. U is called the *search interests terms* for the user.

All terms in the search interests terms for a user form a query Q. Q is applied to the Wikipedia collection to conduct retrieval using the Okapi BM25 model [4]. Each Wikipedia document in the top k ranked results has a *category id*. The k *category id*s form a vector V: <id_1, id_2, ..., id_k>. The vector may contain repeated *category id*s since there may be two or more top ranked Wikipedia documents belonging to the same category. This V is recorded as the user profile based on the category information from Wikipedia.

5. EXPERIMENTAL EVALUATION

In this section we describe an experiment to test our proposed method. In order to do this we need the Wikipedia collection and user logs from a search system. For a search log, we use the data from a Chinese commercial search engine - SOGOU.COM (NASDAQ:SOHU). The data includes one month's user query logs (1.9 GB). The format of each line in the user logs is as follows: $UserId$, $UserQuery$, *RankedPosition*, $RankOfUserClick$, *ClickThroughUrl*.

Each line describes one search activity from one user. $UserId$ is the unique id for this search engine user; $UserQuery$ is a search query input by this user; *RankedPosition* is the ranked position for the click-through URL in the ranked list; $RankOfUserClick$ is the sequence number of the user clicks for this URL; *ClickThroughUrl* is the click-through URL by the user for $UserQuery$. The useful data entries in the user logs for our research are $UserId$, $UserQuery$, and *ClickThroughUrl*.

Table 1 shows an overview of the experimental data. We manually selected 83 users from the SOGOU query logs. For training the user profiles for these users, we have 734 historical search queries with 2311 click-through documents for these queries. Our knowledge system for building user profiles is the simplified Chinese Wikipedia documents. Information about the Chinese Wikipedia collection (dumped in Jan. 2011) [1] is shown in Table 2.

We use the Lemur toolkit [2] to conduct the k-means clustering algorithm on Wikipedia collection, the threshold is set as 0.1 recommended by Lemur. Our Chinese Wikipedia clustering results are shown in Table 3, the distribution of document numbers in categories is shown Figure 2. As shown in Figure 2, about half of the Wikipedia categories have less than 10 documents. This indicates that the clustering algorithm not only groups documents into popular categories, but also places documents into categories with only a few documents. This provides the opportunity to model user search interests at a more specific level. The unsupervised clustering algorithm (k-means) assigns all the Wikipedia documents with a *category* id from 1 to 4785 in this experiment.

For the click-through documents, the top 5 terms ranked by Okapi BM25 weight are treated as the key terms. The Okapi BM25 model is used to search the Wikipedia collection by user search interests terms. The top 10 Wikipedia documents are assumed to be relevant to the user search interests terms. For our results, the generated user interest profiles are vectors consisting of *category id*s. Each *category id* consists of many Wikipedia documents. If a *category id* appears in a user's interest profile frequently, it means this user is very interested in the category. Thus the Wikipedia documents in this category should be very related to the user's search interests. To demonstrate our results, we show examples of some historical user queries and titles of Wikipedia documents in his interested category in Table 4.

6. CONCLUSIONS

In this paper we have described a preliminary study into the utilization of Wikipedia-based category information produced by an unsupervised clustering method to model a user's search interests. The experimental results show that the user's historical queries and click-through documents have the potential to model a user's search interests. This shows that Wikipedia-based personalized modelling is a promising direction to explore for personalized retrieval. Our results successfully create user profiles from the Wikipedia *category id*s. Our future work will focus on utilizing Wikipedia-based user profiles on personalized retrieval tasks.

7. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.

8. REFERENCES

[1] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. CIKM '02, pages 558–565, New York, NY, USA, 2002. ACM.

[2] J. Min, J. Leveling, D. Zhou, and G. J. F. Jones. Document expansion for image retrieval. RIAO '10, pages 65–71, Paris, France, France, 2010.

[3] K. Ramanathan and K. Kapoor. Creating user profiles using wikipedia. In A. Laender, S. Castano, U. Dayal, F. Casati, and J. de Oliveira, editors, *Conceptual*

[1] http://dumps.wikimedia.org/

[2] http://www.lemurproject.org/

Modeling - ER 2009, volume 5829 of *Lecture Notes in Computer Science*, pages 415–427. Springer Berlin / Heidelberg, 2009.

[4] S. Robertson and K. Spärck Jones. Simple, proven approaches to text retrieval. Technical Report UCAM-CL-TR-356, University of Cambridge, Computer Laboratory, Dec. 1994.

[5] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In *CIKM '07*, pages 525–534, New York, NY, USA, 2007. ACM.

[6] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *Proceedings of Workshop on Text Mining, at The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2000)*, 2000.

[7] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. WWW '04, pages 675–684, New York, NY, USA, 2004. ACM.

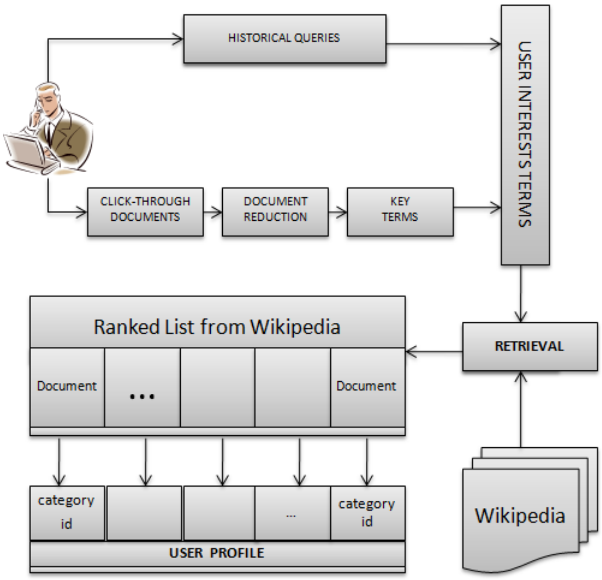
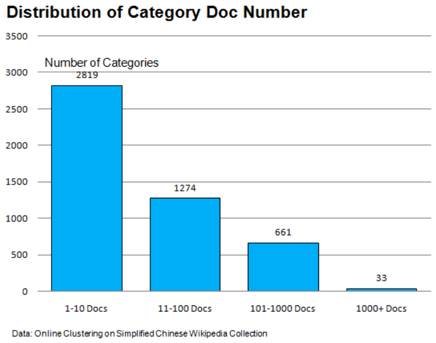


Figure 1: Wikipedia for User Modeling.



Data: Online Clustering on Simplified Chinese Wikipedia Collection

Figure 2: Distribution of Number of Documents in a Cluster.

Table 1: Overview of Experiment Data

Data	Number
Users	83
Training Queries	734
Training Click-Through Links	2311

Table 2: Overview of Chinese Wikipedia Collection

Number of Documents	332,900
Number of Terms	10,959,403
Number of Unique Terms	232,858
Average Document Length	32

Table 3: Results of Wikipedia Clustering

Number of Clusters	4785
Average No. of Documents of a Cluster	70
Largest No. of Documents in a Cluster	15803
Smallest No. of Documents in a Cluster	1

Table 4: Results of User Profiles

User Id	Historical Query	Wikipedia title in category of user profile
user 1	缥缈峰之旅 (name of a novel)	爱情小说 (romantic novel) 虚构角色 (fictional characters)
user 2	台湾报纸 (taiwan newspaper)	日报 (daily newspaper) 台北时报 (a newspaper in taiwan)
user 3	玄幻小说 (mysteries)	安格乌雷尔 (Anglachel)
user 4	电子产品 (electronics)	诺基亚 (nokia) 尼康 (nikon) 佳能 (canon)
user 5	证券 (stock)	证券交易所 (stock exchange) 农贸市场 (farmers market) 证券经纪人 (stock broker) 美国证券交易委员会 (U.S. Securities and Exchange Commission) 资产证券化 (asset securitization) 香港证券经纪业协会 (Hong Kong Stockbrokers Association Limited)