Identifying Vertical Search Intention of Query through Social Tagging Propagation

Ning Liu[1] Jun Yan[1] Weiguo Fan[2] Qiang Yang[3] Zheng Chen[1]

[1]Microsoft Research Asia
Sigma Center, 49 Zhichun Road
Beijing, P.R. China, 100190
{ningl, junyan, zhengc}@microsoft.com

[2]Virginia Polytechnic Institute and State University, Blacksburg, VA, USA 24061-0101
wfan@vt.edu

[3]Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong
qyang@cse.ust.hk

ABSTRACT
A pressing task during the unification process is to identify a user's vertical search intention based on the user's query. In this paper, we propose a novel method to propagate social annotation, which includes user-supplied tag data, to both queries and VSEs for semantically bridging them. Our proposed algorithm consists of three key steps: query annotation, vertical annotation and query intention identification. Our algorithm, referred to as TagQV, verifies that the social tagging can be propagated to represent Web objects such as queries and VSEs besides Web pages. Experiments on real Web search queries demonstrate the effectiveness of TagQV in query intention identification.

Categories and Subject Descriptors
H.3.0 [**INFORMATION STORAGE AND RETRIEVAL**]: *Search process.*

General Terms
Algorithms, Measurement, Performance.

Keywords
Social annotation, metadata, vertical search engine (VSE).

1. INTRODUCTION
Vertical search engines (VSEs) refer to the search services that target at specific information, such as *image*, *video* and *news* search. In recent years, VSEs have become increasingly effective in serving users with specific needs. Unfortunately, many Web users are still unaware of these high quality vertical search resources. Our study in the search query log of a commercial search engine reveals that the number of generic search queries, which have explicit or implicit vertical search intentions, can surpass the traffic of VSEs. This motivates us to develop a unifying approach to bridge user queries and VSEs such that users can see the vertical search results in generic Web search.

In this paper, our solution for this problem is to semantically bridge queries and VSEs by propagating the social annotation, which requires no labeled data for training. With the rapid growth of Web 2.0, a large number of Web users have manually bookmarked their interesting pages through Web platforms such as http://del.icio.us and http://www.digg.com, etc. The user tags for these bookmarks are semantic descriptions of Web pages provided by Web users. However, the abundance of user tags makes us wonder "Besides Web pages, can user tags be leveraged to represent semantics of other Web objects such as queries and VSEs?" With the above motivation, we propose to propagate social annotations, which include user-provided tags of online page bookmarks, to both queries and VSEs, which can then be semantically bridged for query intention identification.

Our novel algorithm, which is called TagQV, functions in three key steps, *query annotation*, *vertical annotation* and *query intention identification*. Query annotation translates queries into tags that are associated with the queries' clicked Web pages. However, the lack of tags for many Web pages leads to the incompleteness of query annotation. We therefore propose a novel approach to automatically tag each Web page by the most frequent terms in associated queries from which users have clicked the page. The vertical annotation step aims to build VSEs' metadata by tags, which associate with pages in their index. The lack of tag information can also make the vertical annotation incomplete. We select the most representative tags of each VSE and expand them to similar ones by calculating similarity among tags. Thus, tags for Web pages are transformed into tags for VSEs. Finally, query intention is identified through the similarity between queries and VSEs in the vector representations of tags. Our experimental results show that TagQV can effectively identify users' intentions for real Web search queries.

WWW 2009, April 20-24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. QUERY INTENTION IDENTIFICATION

2.1 Query Annotation
Nowadays, many Web pages have been tagged by annotators. On the other hand, Web pages that are clicked by users are often good reflection of users' query intentions. Thus bridged by the Web pages, which have both user tags and query clicks, we can semantically connect the search queries and social tags. In this work, we propose to translate queries to social annotation for explaining semantics of short queries. We collect a million pages with full set of tags from del.icio.us. In addition, we also collect 10 days' click-through log of a commercial search engine. Let q, p and t stand for a query, a page and a tag respectively. Figure 1 shows their relationships.

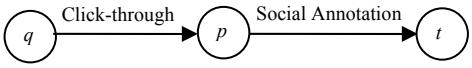

Figure 1. Relationship of query, page and tag

We consider our problem as estimating the probability of a query q_i that can be tagged by tag t_j in the probabilistic model. In other words, we aim to estimate the probability $P(t_j | q_i)$ for query q_i, where i=1,2,…,m is used to distinguish different queries. Given Web pages, we assume queries and tags are conditionally independent. According to this assumption, we have,

$$P(t_j \mid q_i) = \sum_k P(t_j \mid p_k) P(p_k \mid q_i) \quad (1)$$

where p_k , k=1,2,…,h are all pages on the Web. If q_i never clicked p_k, then the probability $P(p_k | q_i)$=0. Thus the summation in Equation (1) only relates to the pages which have been clicked by

q_i. Through this way, the tags which have a high probability to tag a query are used as its metadata. The remaining problem is how to estimate the probability $P(p_k|q_i)$ and $P(t_j|p_k)$ respectively. Note the search click-through log and social annotation make both query-page pairs and page-tag pairs observable. Suppose the number of times that q_i clicks page p_k is f_k. The probability $P(p_k|q_i)$ can be directly approximated by

$$P'(p_k|q_i) = f_k \Big/ \sum_{j=1}^{h} f_j \quad (2)$$

On the other hand, suppose the number of times that p_k is annotated by t_j is g_j, j=1, 2… l, the probability $P(t_j|p_k)$ is approximated by $P'(t_j|p_k) = g_j \Big/ \sum_{i=1}^{l} g_i$ (3)

where t_j, j=1,2,…,l stands for all terms, which can be used as tags.

2.2 Vertical Annotation

Similar to the database selection problem of meta-search engines, another challenge of our problem is how to represent VSEs such that they can be semantically bridged with the queries. Same as the query annotation problem, we aim to estimate the probability of a vertical search engine v_i annotated by tag t_j, i.e. $P(t_j|v_i)$. Ideally, if we can collect all pages p_k , k=1,2,…,h, which are indexed by VSEs v_i, i=1,2,…,m, and collect tags of all these pages, we can approximate two probabilities: $P(p_k|v_i)$, which is the probability of vertical v_i generating page p_k; and $P(t_j|p_k)$, which is the probability of page p_k generating tag t_j. Assume the verticals and tags are conditionally independent if the Web pages are given. Then we have

$$P(t_j|v_i) = \sum_k P(t_j|p_k)P(p_k|v_i) \quad (4)$$

The probability $P(t_j|p_k)$ can be approximated by Equation (3), which is the same as query annotation. Thus the remaining problem is how to estimate $P(p_k|v_i)$. Since it is hard to collect all Web pages which are indexed by the VSEs, in this work we propose to use the pages which have been clicked in the VSEs as an approximation. Without loss of generality, let p_k , k=1,2,…,h stands for all the Web pages which have been clicked in any VSE. Given a vertical v_i and its vertical search click-through log, suppose p_k was clicked in v_i for r_k times, then we have,

$$P(p_k|v_i) = r_k \Big/ \sum_{j=1}^{h} r_j \quad (5)$$

Thus Equation (4) can be computed by Equations (3) and (5).

2.3 Query Intention Identification

Bridged by social annotation, we can compute the similarity between a query and a VSE. Then the vertical search intention of queries can be identified by similarities. As an intuitive example, in our 10 days' click-through log of video search, about 10% of the clicked pages have been tagged as "YouTube" by users. Thus "Youtube" has a high score in the *video* search. If the tag of a query also includes the tag "YouTube" with a high probability, then this query is relevant to the *video* search. In reality, the similarity between a query and a vertical search engine is not bridged by one unique tag. Thus given a query q_j and a vertical v_i, suppose the conjunction of their corresponding tags are t_k, k=1,2,…,n, we define the similarity between q_j and v_i by the Cosine similarity in vector space of tags,

$$I(q_j, v_i) = \sum_{k=1}^{n} p(t_k|q_j)S_i(t_k) \Big/ \sqrt{\sum_{l=1}^{n} p(t_l|q_j)^2} \sqrt{\sum_{l=1}^{n} S_i(t_l)^2} \quad (6)$$

where we define $S_i(t_k)$=0 if t_k is not in the tag list of v_i. Similarly, $P(t_k|q_j)$=0 if t_k is not a tag of q_j. Finally, we identify the query intention by threshold η , i.e., if $I(q_j, v_i) \geq \eta$, we identify query q_j has the intention to search in vertical v_i. The relevance of the verticals to query q_j is ranked by the similarity in Equation (6).

3. EXPERIMENTS

In this section, we use real log data to help show why we need to identify vertical search intentions of generic Web search queries. We utilize a 10 days' vertical search query log and a Web search query log of a commonly used commercial search engine. And then we randomly sample queries from the beginning to the end of this list to guarantee that we have both high- and low-frequency queries. We ask some labelers to label these 3,000 queries. Given a query, a user label its intention with 5 different scores from 0 to 4, where 4 stands for "strongly related to a vertical" and 0 stands for "not related".In Figure 1, we show Precision, Recall, F-measure[3] of TagQV where s the baselines are "QC"[1], "Meta" [2] and the results from Google.

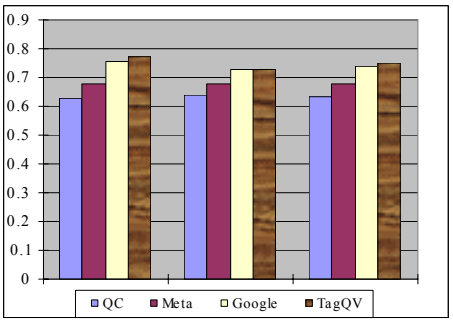

Figure 1. Evaluate different approaches

These results tell us that TagQV can perform better than some classical meta-search and query classification strategies in the query intention identification task.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed to semantically bridge generic Web search queries with vertical search engines through social annotation. We showed that the social annotation of Web pages can also be used to annotate other Web objects such as queries and vertical search engines. Through propagating the social annotation, these Web objects can be effectively connected. In addition, to solve the incompleteness of social annotation, we propose to automatically tag Web page through click-through information of queries. Experimental results show that our proposed TagQV algorithm can better identify users' search intentions than some other baseline approaches. Through applying the query annotation part of TagQV algorithm, we propose a list of potential VSEs which we believe will be interested by search users, though some of them have already existed.

5. REFERENCES

[1] Meng, W.Y., Yu, C. and Liu, K.L. Building efficient and effective metasearch engines. ACM Computing Surveys (March 2002), 34(1), 48-89.

[2] Shen, D., Pan, R., Sun, J.T., Pan, J.J., Wu, K.H., Yin, J. and Yang, Q. Query enrichment for web-query classification. ACM TOIS (July 2006), 24 (3), 320-352.

[3] Yang, Y.M. An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval (1999), 1 (1/2), 67-88.