Learning-Based Time-Sensitive Re-Ranking for Web Search

Po-Tzu Chang, Yen-Chieh Huang, Cheng-Lun Yang, Shou-De Lin, Pu-Jen Cheng
Graduate Institute of Computer Science and Information Engineering
Graduate Institute of Network and Multimedia
National Taiwan University
{r98922102, r98944011, r99944042, sdlin, pjcheng}@csie.ntu.edu.tw

ABSTRACT

To model time-dependent user intent for Web search, this paper proposes a novel method using machine learning techniques to exploit temporal features for effective time-sensitive search result re-ranking. We propose models to incorporate users' click through information for queries that are seen in the training data, and then further extend the model to deal with unseen queries considering the relationship between queries. Experiment shows significant improvement on search result ranking over original search outputs.

Categories and Subject Descriptors

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

General Terms: Algorithms, Design

Keywords: Time sensitive query, temporal features

1. INTRODUCTION

Traditional search engine uses keyword matching with link-based strategies, such as PageRank, to improve search results without considering users' intent. We have observed through web query logs that the ideal ranking of search results does depend on query intent, which usually vary with time. For the same query, the users may want to find different web pages in different time of the day. For example, from AOL dataset, we found that users who enter the query term "aa" prefer visiting substance control website (www.alcoholics-anonymous.org) from 7am to 2pm, but the visiting preference changes to American Airline website (www.aa.com) from 3pm to 3am next day (see Figure 1). The aim of this paper is to re-rank search results based on user intents at different time of a day.

2. RELATED WORK

Several researchers have observed that query popularity changes over time. Beitzel et al. [7] show changes in popularity of topically categorized queries across different hours of a day. Anagha et al. [6] explore how queries, their associated documents, and the query intent change over the course of 10 weeks by analyzing query log data. Dong et al. [2] propose ideas of "recency ranking", which refers to ranking documents by relevance which takes freshness into account. They use multiple recency features to provide temporal evidence which effectively represents document recency. Li and Croft [4] identify a type of query that favors very recent documents, and propose a time-based language model to retrieve these queries. However, above works do not focus on exploiting queries intent with time to improve the search quality.

3. METHODOLOGY

We use the users' temporal click information from query logs as training data to re-rank the search results at different time period. We design models to handle queries with click history, *existing queries (see 3.1)*, and queries with few or no click information, *missing queries (see 3.2)*. Finally, we use a learning-based method to build an ensemble of our predictions for search result re-ranking. The training data contains the aggregated user click counts, with time stamp, of top URLs returned for each query. We use several different models to generate a raw score for each <query, URL> pair at each time period. Then, we exploit a learning approach to combine the scores from different models and rank each URL for each given query. Finally, such ranking is linearly combined with the original (or baseline) ranking results through the following re-ranking formula, $Score(u) = \frac{\alpha}{original\_rank(u)} + \frac{(1-\alpha)}{predicted\_rank(u)}$, *where α is a weighting parameter varied from 0.0 to 1.0.*

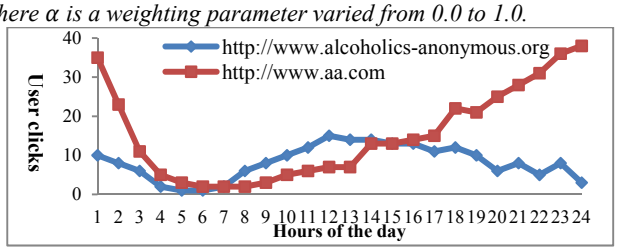

Figure 1. Example of time-sensitive query. For the same query, "aa", user preference changes during the day.

3.1 Models for Existing Queries

Given the query and time, we propose several models to estimate the quality of each returned documents for the re-ranking task.

Temporal Prior: The relevance of a document/URL u to an existing query q issued at time t is defined as: $P(u|q,t) = \frac{f_t(q,u)}{\sum_{\forall u_i} f_t(q,u_i)}$, where $f_t(q,u)$ is the frequency of users visiting URL u when searching query q at time t. This stands for u's prior clicking frequency probability given query q at time t in training data. This can be considered as a maximum-likelihood model.

Time-sensitive Content Model: Previous model only works for URLs that has been seen in the training set. To overcome such limitation, we propose to exploit the similarity between unseen documents. The relevance of a URL u to query q issued at time t is defined as: $P(u|q,t) \propto \sum_{\forall q'} Sim(u,u') * P_{click}(u'|q,t)$ where u' represents the URL that are already in the training corpus and $Sim(u,u')$ is the context similarity of u and u'. Sim(u,u') can be estimated from comparing the difference between the n-gram distributions generated from u and u', and $P(u'|q,t)$ is the temporal prior of u'.

Temporal PageRank: PageRank is a popular model to assess the authority of a page. Unfortunately, the original PageRank algorithm does not consider temporal information. Here we propose temporal PageRank to incorporate such information by adding time information (click count of each URL u at time t) into PageRank. To facilitate such idea, we first build a directed graph of all URLs in the dataset. There is a link connecting two webpages if the content of one URL contains a hyperlink to another URL. Then, we assign a click number for each webpage which is the total number of user clicks for that URL at time t. Temporal PageRank function is defined below.

$$R(t) = \begin{bmatrix} (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} \begin{bmatrix} P(u_1|U,t) \\ \vdots \\ P(u_n|U,t) \end{bmatrix} + d \begin{bmatrix} l(u_1,u_1) & \dots & l(u_1,u_n) \\ \vdots & \ddots & \vdots \\ l(u_1,u_n) & \dots & l(u_n,u_n) \end{bmatrix}$$

$l(u_i,u_j)$ is the adjacency function of these URLs, N is the total number of URLs, and d is the damping factor. The main difference from the original PageRank equation lies in $P(u_i|t) = \frac{f(u_i,t)}{f_{\forall u_i \in U} \max(u_i,t)}$, which represent the click opportunity of URL u at time t. Note that this model measures the temporal authority of pages, and thus is query independent (similar to PageRank).

SIGIR'12, August 12-16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

3.2 Dealing with Missing Queries

Missing queries presents a more difficult challenge because of the insufficient temporal click information to learn from. To solve this problem, we propose to exploit click information from *extended queries* which we have sufficient click information. First, we build a bi-partite graph of all queries and URLs as shown in Figure 3. We eliminate popular hub URLs such as Amazon.com, to avoid the interference from hub sites. Then, we define a query set Q', which contains queries q' that are 2 steps away from missing query q, as the extended queries, and generate new $f_t(q,u)$ by combining user clicks of q and q' as defined in the following equation, $f_t'(q,u) = f_t(q,u) + \sum_{q' \in Q'} f_t(q',u)$. Note that $f_t(q,u)$ is either 0 or a very small number based on the definition of missing query.

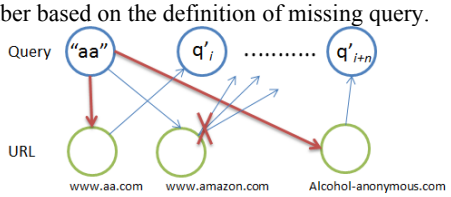

Figure 3. Query Expansion after removing hub URLs

Since extended queries may have different degree of resemblance to the missing query, we propose to weight click counts based on the similarity between them. We define two similarity measures:

1. $\mathrm{Sim}(q,q') = \sum_{u \in U, u' \in U'} \mathrm{LM}(u,u')$ where $\mathrm{LM}(u,u')$ is the content similarity of u and u' using character-based Language Model described in [10] while U and U' are the set of URLs returned by search engine for query q and q' respectively.
2. We further consider the click count of users using: $\mathrm{Sim}_{wgt}(q,q') = \sum_{u \in U, u' \in U'} \mathrm{LM}(u,u') * P_{\mathrm{click}}(u'|q')$.

We use the similarity measures defined in this section as features in Support Vector Regression for missing query dataset.

3.3 Model-Ensemble through Regression

Eventually we want to combine the results from proposed models with global features, including query term, query time, click count, and target URL. Instead of manually choosing the ensemble weights, we propose to learn them through a regression model. We choose epsilon-SVR with linear kernel for its superior performance.

4. EXPERIMENTS

We collect data from the AOL query log, which contains click-through records from March 1, 2006 to May 31, 2006, representing the search behavior of 650,000 users generating 20 million queries. Each record has five fields, user id, query, query time, URL, and the relevance rank of each URL. The relevance rank is the default ranking of URLs for a query given by AOL search engine. Note that here we assume the query time in AOL dataset reflects the local time zone instead of an adjusted global time. To verify our hypothesis, we compare it with the query log from Sogou, a popular search engine in China, as China uses one time zone throughout the country. The result shows similar searching behavior over time (see Figure 4), and support our hypothesis that AOL data reports local time zone, where there are fewer records from midnight to early morning. If the time were converted to a global time zone in AOL, it is likely we would have seen more evenly distributed graph throughout the day.

We merge the clicks within four-hour interval starting from 2am, resulting in six time periods. Then 70% raw data are selected at random as training with the remaining 30% as testing and the process repeats 10 times. Existing queries are defined as queries containing two or more URLs whose cumulative click counts are greater than 50 in the training data. Missing queries are those containing two or more URLs whose cumulative click counts are greater than 25 in testing data, but do not appear in the training data. The baseline is determined by the original search ranking given in the AOL query log. For ground truth, we use the click number rankings from testing data. We calculate the Kendall's tau value between the ground truth and the ranking we predicted. Note that the baseline (i.e. without re-ranking, $\alpha = 1$) reaches Kendall's tau value 0.570.

Existing Queries: The first row of Table 1 shows a decent improvement of 9.1% on content model over the baseline. Note that the temporal prior model is not listed as it cannot be generated with missing URLs. The final row of Table 1 shows using all features achieves the best result of 10.3% over the baseline when α=0.2.

Table 1. Kendall's tau of proposed models.

α	0	0.2	0.5	0.7	0.9
Time-sensitive Content Model	0.658	0.661	0.629	0.589	0.574
Temporal PageRank	0.594	0.607	0.618	0.596	0.585
SVR w/Time-sensitive Content Model	0.657	0.661	0.626	0.590	0.573
SVR w/ All Models	0.669	0.673	0.634	0.590	0.573

Missing Queries: On top of the global features, our learner adds similarity scores obtained from Section 3.2 for ensemble. The Kendall's tau value of the original order is 0.189, much lower than that of existing queries. The baseline is not as good because these queries usually contain typos or unpopular combinations of terms, which causes problems for a search engine due to lack of training data. Our model achieves Kendall's tau value of 0.309, a significant improvement of 12%. The results show that considering users' temporal intent can bring relatively more information to missing queries than to existing ones.

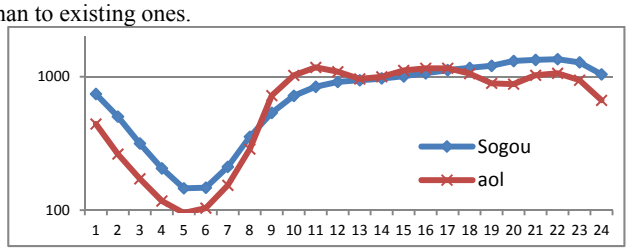

Figure 4. Number of searches by the hour. X-axis is hour of the day. Y-axis is log (number of searches)/1000.

5. CONCLUSION

This study verifies that understanding how users' search intents change over time is critical for a search system. The consistent improvement with a wide range of α, together with the fact that better improvement can be achieved when α decreases (i.e. the weight of our model increases) show the effectiveness of this framework. Future work includes exploiting advanced models (e.g. probabilistic graphical model) and design more temporal features for learning.

6. ACKNOWLEDGEMENT

This work was supported by National Science Council, National Taiwan University and Intel Corporation under Grants NSC 100-2911-I-002-001, and 101R7501.

REFERENCES

[1] Elsas, J. and Dumais, S. T. 2010. Leveraging temporal dynamics of document content in relevance ranking. In *Proc. of ACM WSDM Conference*, 2010.

[2] Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Buchner, K., Zhang, R., Liao, C. and Diaz, F. 2010. Towards recency ranking in Web search. In *Proc. of ACM WSDM*, 2010.

[3] Kulkarni, A., Teevan, J., Svore, K., and Dumais, S. Under-standing temporal query dynamics. In *Proc. of WSDM*, 2011.

[4] Li, X. and Croft, W. B. 2003, Time-based language models. In *Proc. of ACM CIKM Conference*, 2003.

[5] Zhang, R., Chang, Y., Zheng, Z., Metzler, D. and Nie, J.-Y. 2009. Search result re-ranking by feedback control adjustment for time-sensitive query. In *Proc. of NAACL*, 2009.

[6] Anagha K., Jaime T., Krysta M. S. and Susan T. D. 2011. Temporal Query Dynamics. In *Proc. of ACM WSDM*, 2011.

[7] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. and Frieder, O. 2004. Hourly analysis of a very large topically categorized Web query log. *In Proceedings of SIGIR 2004*.

[8] Zhao, Q., Hoi, C.H., Liu, T.Y., Bhowmick, S.S, Lyu, M.R., Ma, W.Y. 2006. Time-Dependent Semantic Similarity Measure of Queries Using Historical Click-Through Data. In *Proc. of WWW,* 2006.

[9] Baraglia, R., Castillo, C., Donato, D., Nardini, F.M., Perego, R. 2010. The Effects of Time on Query Flow Graph-based Models for Query Suggestion. *In Proceedings of RIAO*.

[10] Carpenter, B. 2005. Scaling high-order character language models to gigabytes. *Proceedings of the 2005 Association for Computational Linguistics Software Workshop*.