Enhancing Topical Ranking with Preferences from Click-Through Data

Yi Chang, Anlei Dong, Ciya Liao, Zhaohui Zheng
Yahoo! Labs
701 First Avenue
Sunnyvale, CA 94089
yichang, anlei, ciyaliao, zhaohui@yahoo-inc.com

ABSTRACT

To overcome the training data insufficiency problem for dedicated model in topical ranking, this paper proposes to utilize click-through data to improve learning. The efficacy of click-through data is explored under the framework of preference learning. The empirical experiment on a commercial search engine shows that, the model trained with the dedicated labeled data combined with skip-next preferences could beat the baseline model and the generic model in NDCG_5 for 4.9% and 2.4% respectively.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*Relevance Feedback*; H.4.m [**Information Systems**]: Miscellaneous—*Machine learning*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

topical ranking, preference learning, click-through data

1. INTRODUCTION

Topical ranking is to rank documents for a specific topic, which is usually for a category of queries, such as *navigational* queries, *news* queries, *product* queries, etc. A topical ranking needs a dedicated model for the queries belonging to this category (topic); such a divide-and-conquer strategy is different from a generic model for the ranking of all queries [2]. The amount of training data dedicated to one topic is usually insufficient because human labeling is expensive and time-consuming. We propose to extract click-through data and incorporate it with dedicated training data to learn dedicated model. Specifically, pair-wise preference data, including skip-above pairs and skip-next pairs, is exploited. From the aspect of learning algorithm, we adopt GBrank algorithm [3] to learn the ranking model because GBrank algorithm has proved to be one of the most effective up-to-date learning-to-rank algorithms; furthermore, GBrank algorithm also takes preference pairs as inputs.

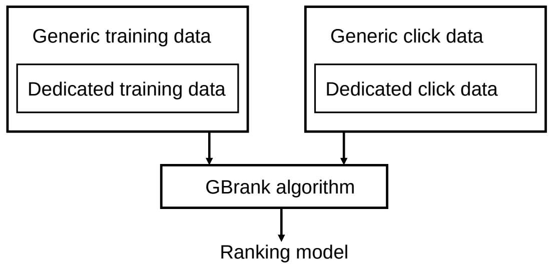

Figure 1: framework of incorporating click-through data with training data to improve dedicated model for topical ranking

The contributions of this paper are: 1) exploitation of click-through data to improve dedicated model for topical ranking; 2) exploration of the factors of click-through data in helping dedicated model learning.

2. APPROACHES AND EXPERIMENTS

Figure 1 illustrates the framework of incorporating click-through data with training data to improve dedicated model for topical ranking, for which we seek the best way to utilize click-through data.

We use heuristic rules to extract skip-above pairs and skip-next pairs, which are similar to Strategy 1 (click > skip above) and Strategy 5 (click > no-click next) proposed in [1]. To reduce the misleading effect of an individual click behavior, click information from different query sessions is aggregated before applying heuristic rules. For a tuple $(q, \mathrm{url}_1, \mathrm{url}_2, \mathrm{pos}_1, \mathrm{pos}_2)$ where q is query, url_1 and url_2 are urls representing two documents, pos_1 and pos_2 are ranking positions for the two documents with $\mathrm{pos}_1 \prec \mathrm{pos}_2$ meaning url_1 has higher rank than url_2, the statistics for this tuple are listed in Table 1.

Skip-above pair extraction: if ncc is much larger than cnc, and $\frac{cc}{imp}$, $\frac{ncnc}{imp}$ is much smaller than 1, that means, when url_1 is ranked higher than url_2 in query q, most users click url_2 but not click url_1. In this case, we extract a skip-above pair, i.e., url_2 is more relevant than url_1. In order to have highly accurate skip-above pairs, a set of thresholds are applied to only extract the pairs that have high impression and ncc is larger enough than cnc.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

Table 1: Statistics of click occurrence

imp	impression, number of occurrence of the tuple
cc	number of occurrence of the tuple where two documents both get clicked
ncc	number of occurrence of the tuple where url_1 is not clicked but url_2 is clicked
cnc	number of occurrence of the tuple where url_1 is clicked but url_2 is not clicked
ncnc	number of occurrence of the tuple where url_1 and url_2 are not clicked

Table 2: Experiment data

Generic training data	4,429 queries, 121,061 q-urls
Dedicated training data	757 queries, 18,174 q-urls
Dedicated testing data	75 queries, 1,857 q-urls
Generic skip-above	0.55M preferences
Generic skip-next	0.7M preferences
Dedicated skip-above	29,723 preferences
Dedicated skip-next	279,408 preferences

Skip-next pair extraction: if $pos_1 = pos_2 - 1$, cnc is much larger than ncc, and $\frac{cc}{imp}$, $\frac{ncnc}{imp}$ is much smaller than 1, that means, in most of cases when url_2 is ranked just below url_1 in query q, most users click url_1 but not click url_2. In this case, we regard this tuple as a skip-next pair.

Experiments: We do experiments using the data obtained from a commercial search engine, including training data, testing data and click-through data, which are called generic data. We apply a query classifier to detect dedicated data from generic data. The data details are described in Table 2. We use $NDCG_5$ to evaluate ranking model.

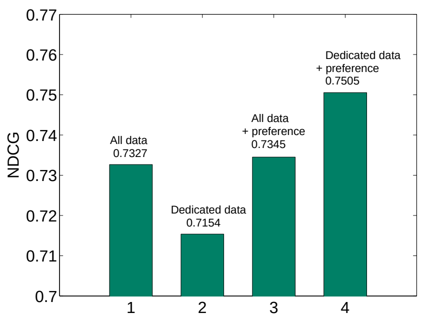

Figure 2: NDCG comparison with different training data.

The experiments are designed to address two questions: 1) can a dedicated model outperform generic model on the predefined query category? 2) what is the empirical results using skip-above or skip-next preferences respectively?

Figure 2 demonstrates $NDCG_5$ comparison over different training data. Due to insufficiency of dedicated labeled data which is only 15% of the generic labeled data, the generic model is more than 2% better than the dedicated model in $NDCG_5$. Another reason is data correlation among different query categories: although a larger portion of labeled data does not belong to the predefined category, they are also useful for training the dedicated model because both the dedicated and the generic model would share some common structure or common pattern. Keep them in the training data could improve generalization capability of a ranking model.

After adding click-through preferences extracted from click-through data, $NDCG_5$ observation is opposite: click-through based preferences only contributes 0.25% $NDCG_5$ improvement over the generic training data; however, combining the dedicated training data and the click-through based preferences together could generate the best model, which show 4.91% $NDCG_5$ improvement over the dedicated model trained only with the dedicated labeled data, and it is also 2.44% over the generic model. In general, a dedicated model does outperform the generic model on the predefined query category. Those click-through based preferences under the query category do provide some novel information to improve the ranking.

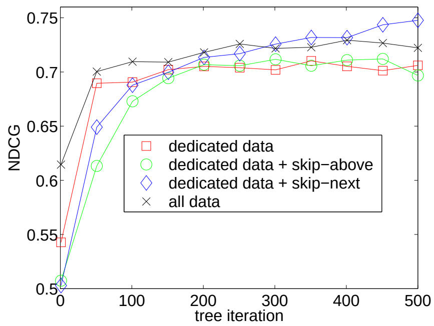

Figure 3: NDCG comparison with different preferences.

Figure 3 shows $NDCG_5$ trends over 500 tree iterations of different models. We observe that using skip-next preferences yields better model than using skip-next preferences. There are two reasons to explains this: 1) Essentially, skip-next preferences are consistent with baseline model while skip-above preferences are inconsistent with baseline model. As the dedicated training data is insufficient, the utility of extra consistent data is higher than the utility of extra inconsistent data. 2) There are 18% skip-above preferences which are inconsistent with human labeling, while there are only 4% skip-next preferences which are inconsistent with human labeling. The user click could be easily disturbed by many factors, such as snippet display, word highlight. As NDCG measurement is based on human labeling, the utility of skip-above pairs is hurt due to the high inconsistency with human labeling.

3. CONCLUSIONS

This paper has demonstrated the efficacy of utilizing click-through data to improve dedicated model learning for topical learning. There are quite a few promising directions along this research work, such as reduce the inconsistency between skip-above preferences and human labeling.

4. REFERENCES

[1] T. Joachims, L. Granka, B. Pan, and G. Gay. Accurately interpreting clickthough data as implicit feedback. *Proc. of ACM SIGIR Conference*, 2005.

[2] T. Y. Liu. Learning to rank for information retrieval. *SIGIR tutorial*, 2008.

[3] Z. Zheng, K. Chen, G. Sun, and H. Zha. A regression framework for learning ranking functions using relative relevance judgments. *Proc. of ACM SIGIR Conference*, pages 287–294, 2007.