A Probabilistic Model based Approach for Blended Search

Ning Liu, Jun Yan, and Zheng Chen
Microsoft Research Asia
Sigma Center, 49 Zhichun Road
Beijing, P.R. China
ningl@microsoft.com, junyan@microsoft.com, zhengc@microsoft.com

ABSTRACT
In this paper, we propose to model the blended search problem by assuming conditional dependencies among queries, VSEs and search results. The probability distributions of this model are learned from search engine query log through unigram language model. Our experimental exploration shows that, (1) a large number of queries in generic Web search have vertical search intentions; and (2) our proposed algorithm can effectively blend vertical search results into generic Web search, which can improve the Mean Average Precision (MAP) by as much as 16% compared to traditional Web search without blending.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval – *Retrieval models*

General Terms
Algorithms, Verification.

Keywords
Vertical search, blended search, language model, query log.

1. INTRODUCTION
Vertical Search Engine (VSE) refers to the search services that target at specific kind of 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. For example, many users enter "Jennifer Lopez image" in generic Web search to find images of Jennifer Lopez instead of entering "Jennifer Lopez" in Image search. Our study in the search query log of a commercial search engine reveals that the number of generic Web search queries, which have explicit or implicit vertical search intentions, can surpass the traffic of VSEs. As more evidences, we found that 12.3% of generic Web search queries have strong *image* search intention and 13.1% have the *news* search intention.

To make the value of VSEs aware by search users, many efforts have been made by commercial search engines. Yahoo Shortcut, Ask 3D, MSN Live and Google Universal Search, are all examples of these efforts. However, why and how to blend the vertical search results into generic Web search are still under explored in academia. In this paper, we propose a probabilistic model based approach as a solution to the blended search and show that the blended search can truly help general purpose search engines. As related work, the "blended search" can be naturally considered as a meta-search problem [4]. For example, if we treat both generic Web Search and VSEs as component search engines, the blended search is to aggregate the search results from these components into a single list. However, from the classical meta-search problem's configuration, the query log of component search engines is not available for study.

In this extended abstract, we model the blended search problem based on the conditional dependencies among queries, VSEs and all the search results. We utilize the usage information, i.e. query log, of all the VSEs, which are not available for traditional meta-search engines, to learn the model parameters by the smoothed unigram language model. Finally, given a user query, the search results from both generic Web search and different VSEs are ranked together by inferring their probabilities of relevance to the given query. The main contributions of this work are, (1) through studying the belonging vertical search engines' query log of a commercial search engine, we show the importance of blended search problem; (2) we propose a novel probabilistic model based approach to explore the blended search problem; and (3) we experimentally verify that our proposed algorithm can effectively blend vertical search results into generic Web search, which can improve the MAP as much as 16% in contrast to traditional Web search without vertical search blending and 10% to some other some ranking baseline.

2. MOTIVATIONAL OBSERVATIONS
In this section, we show why we study blended search. We utilize 10 days' vertical search query log and a general Web search query log of a commercial search engine as our dataset. We study five VSEs, which are *Image*, *Video*, *News*, *Books* and *RSS Feeds*. If only considering the English queries, the traffic of the five VSEs is only 1.56% of the Web search traffic. We assume that if a VSE's name v appears as part of a query q in two special patterns, then q has the explicit intention to search in VSE v. The two patterns are "*v of* something" and "something *v*". For example, the queries "*image of* Britney Spears" and "Britney Spears *image*" are assumed to have the explicit image search intentions. Under this assumption, there are 2,766,937 queries (about 1%) having the explicit vertical search intentions to the 5 VSEs in the 10 days' search query log. We can extent the vertical search engine name v to more related terms in the above assumption. Take the image search VSE as example. If we allow v to be any term in "*image, picture, gallery, wallpaper* and *pic*", there are 2,573,911 queries. Note the traffic of image search is 2,106,452. The number of Web search queries that have explicit image search intentions, have already surpassed the traffic of image search.

To do a more accurate study, we randomly sampled 2,153 generic Web search queries and ask labelers to judge the relevance between queries and search domains, which are Image, Video, News, RSS and Book. We use 4-level labeling with "3" is "strong relevant" and "0" is "irrelevant". The results are given in Table 1 It is interesting that 12.3% queries have strong image search intention, 8.5% have strong video search intention and 13.1% have strong news search intention. These statistics motivated us to study the blended search.

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

Table 1. Vertical intention of generic Web-search queries

	Image	Video	News	Blog	Book
3	**12.3%**	8.5%	13.1%	0.9%	7.2%
2	26.1%	15%	28%	8.6%	11.9%
1	36.3%	46.6%	49.5%	70.7%	29.9%
0	25.7%	29.7%	9.3%	19.7%	50.9%

3. TYPESET TEXT

In this paper, search engines (SEs) is noted as both VSEs and general purpose search engine. Let $\Lambda = \{S_i, i=1,2,\ldots m\}$ stands for the set of all SEs. Given the user query q, the search results are represented by $R = \{R_{ij}, i=1,2,\ldots m$ and $j=1,2,\ldots n\}$, where R_{ij} is the j^{th} search result of query q in the i^{th} search engines Si. Our problem is to rank the results come from various SEs for generic Web search. Thus we aim to learn the probability $p(R_{ij}|q)$ and then the search results of q can be ranked according to the probabilities in a decreasing order. To learn $p(R_{ij}|q)$, we firstly analyze the dependencies between queries, SEs and search results. Since R_{ij} is retrieved from Si by query q, it depends on both Si and q. On the other hand, we assume that R_{ij} is independent of V_l if $l \neq i$. Intuitively, we assume that each result can only be retrieved from one search engine. Note the fact that different queries will have different vertical search intentions. The importance of SEs is not uniformly distributed. It will highly depend on the user query. Motivated by these clear dependencies, we modeled our problem, i.e. the problem of learning probability $p(R_{ij}|q)$, by a probabilistic graphical model, which is described by Figure 1.

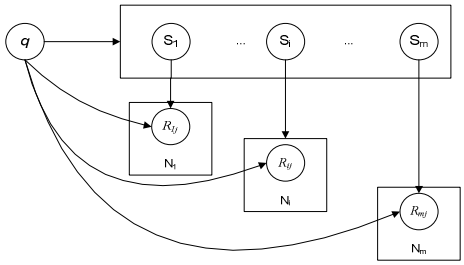

Figure 1. The proposed model for blended search.

According to the conditional independency assumptions, we can get the probability distribution $p(R_{ij}|q)$ through,

$$\begin{aligned} p(R_{ij} \mid q) &= \sum_l p(R_{ij}, S_l \mid q) \\ &= \sum_l p(R_{ij} \mid S_l, q) p(S_l \mid q) \\ &= p(R_{ij} \mid S_i, q) p(S_i \mid q) \end{aligned} \quad (3.1)$$

The remaining problem is to estimate $p(R_{ij} \mid S_i, q)$ and $p(S_i \mid q)$. Through Bayesian $p(S_i \mid q) \propto p(q \mid S_i) p(S_i)$, where $p(S_i)$ is our prior belief that Si relates to any arbitrary query q and $p(q \mid S_i)$ is known as the query likelihood. Without loss of generality, suppose query q consists of k terms $\{t_1, t_2, \ldots t_k\}$. We utilize the unigram language model, which is a multinomial model, to estimate the probability $p(q \mid S_i)$ [1]. On the other hand, we propose to estimate the probability distributions $p(R_{ij} \mid S_i, q)$ by,

$$p(R_{ij} \mid S_i, q) = \frac{Score(R_{ij} \mid S_i, q) \cdot e^{-\lambda N_j}}{\sum_l Score(R_{ij} \mid S_i, q) \cdot e^{-\lambda N_l}} \quad (3.2)$$

where $e^{-\lambda N_j}$ is a factor for punishing the vertical search results and N_j is the position of I_{ij} in the result list. The intuition for the punishing factor is that we embed the vertical search results into generic Web search if and only if we are confident they should be embedded, i.e. they are very relevant among all results. The scoring function $Score(R_{ij} \mid S_i, q)$ is the relevance score of R_{ij} from S_i for query q, which should be provided by search engines. As an approximation, in this work, we define

$$Score(R_{ij} \mid S_i, q) = \exp\{-r_i\}$$

where r_i is the ranked position of R_{ij} in S_i if it is returned by S_i.

4. EXPERIMENTS

The dataset used for experiments is the 2,153 labeled queries, which is introduced in section 2. For each of these queries, its top 10 vertical search results and top 10 generic Web search results are mixed together for the labelers to rank their relevance. A voting strategy is used to determine the data ground truth. The evaluation metrics are Mean Average Precision (MAP) and Normalized Discount Cumulative Gain (NDCG) [3]. For comparison purpose, three baseline strategies are involved to compare with our proposed algorithm. The first is a learning-based meta-search method, called ProFusion [2] algorithm. The second baseline is Random Blended which merges the top results from different SEs randomly. We also use generic Web search we crawled from the commercial search engine we used as the baseline. The results are shown in table 2. PM stands for our proposed approach and RB is the random results. Table 2 verifies the significant improvement of our proposed approach.

Table 2. Comparison with other methods

	PM	ProFusion	RB	Web
MAP	**0.82**	0.6912	0.5218	0.732
NDCG	**0.705**	0.6566	0.3945	0.6613

5. CONCLUSION

Vertical search engines (VSEs) have attracted much attention in the past decade. However, they are not very popular yet. This motivates the commercial search engines to blend vertical search results into generic Web search. In this abstract, we first study the vertical search engines' query log of a commercial search engine to show the importance of blended search problem. And then we propose a probabilistic model based approach to explore the blended search problem. Finally we experimentally verify that our proposed algorithm can effectively blend vertical search results into generic Web search.

6. REFERENCES

[1] Borthwick, A. "Survey Paper on Statistical Language Modeling", Technical Report, Proteus project, New York University Computer Science Department, 1997

[2] Gauch, S., Wang, G. and Gomez, M. Profusion: intelligent fusion from multiple, distributed search engines. J. Univers. Comput. Sci. 2(9), (1996), 637-649.

[3] Järvelin, K. and Kekalainen, J. IR evaluation methods for retrieving highly relevant documents. In Proceedings of the *23rd* annual international ACM SIGIR conference (Athens, Greece, July, 2000) SIGIR'00, ACM Press, New York, NY, 41-48.

[4] Meng, W., Yu, C., and Liu, K. Building efficient and effective metasearch engines. ACM Comput. 34(1) 48-89.