Ranking with Query-Dependent Loss for Web Search

Jiang Bian*
College of Computing
Georgia Institute of Technology
jbian@cc.gatech.edu

Tie-Yan Liu, Tao Qin
Microsoft Research Asia
{tyliu, taoqin}@microsoft.com

Hongyuan Zha
College of Computing
Georgia Institute of Technology
zha@cc.gatech.edu

ABSTRACT

Queries describe the users' search intent and therefore they play an essential role in the context of ranking for information retrieval and Web search. However, most of existing approaches for ranking do not explicitly take into consideration the fact that queries vary significantly along several dimensions and entail different treatments regarding the ranking models. In this paper, we propose to incorporate query difference into ranking by introducing query-dependent loss functions. In the context of Web search, query difference is usually represented as different query categories; and, queries are usually classified according to search intent such as navigational, informational and transactional queries. Based on the observation that such kind of query categorization has high correlation with the user's different expectation on the result accuracy on different rank positions, we develop position-sensitive query-dependent loss functions exploring such kind of query categorization. Beyond the simple learning method that builds ranking functions with pre-defined query categorization, we further propose a new method that learns both ranking functions and query categorization *simultaneously*. We apply the query-dependent loss functions to two particular ranking algorithms, RankNet and ListMLE. Experimental results demonstrate that query-dependent loss functions can be exploited to significantly improve the accuracy of learned ranking functions. We also show that the ranking function jointly learned with query categorization can achieve better performance than that learned with pre-defined query categorization. Finally, we provide analysis and conduct additional experiments to gain deeper understanding on the advantages of ranking with query-dependent loss functions over other query-dependent ranking approaches and query-independent approaches.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*Retrieval models*

*The work was done when the first author was intern at Microsoft Research Asia

WSDM'10, February 4–6, 2010, New York City, New York, USA.

General Terms

Algorithms, Experimentation, Theory

Keywords

Ranking for Web search, query-dependent loss function, query difference

1. INTRODUCTION

Ranking has become an essential research issue for informational retrieval and Web search, since the quality of a search system is mainly evaluated by the relevance of its ranking results. The task of ranking in the search process can be briefly described as follows. Given a query, the deployed ranking model measures the relevance of each document to the query, sorts all documents based on their relevance scores, and presents a list of top-ranked ones to the user. Thus, the key problem of search technology is to develop a ranking model that can best represent relevance.

Many models have been proposed for ranking, including the Boolean model [1], vector space model [20], probabilistic model [18] and language model [13, 17]. Recently, there are renewed interests in exploring machine learning methodologies for building ranking models. Many learning-based approaches have been introduced, some popular examples of which include MCRank [15], RankNet [5], RankSVM [11], RankBoost [8], GBRank [26], ListNet [6], ListMLE [23], and IsoRank [27]. These approaches leverage training data, which consists of queries with their associated documents and relevance labels, and machine learning techniques to make the tuning of ranking models theoretically sound and practically effective.

In most of the previous works, the significant difference in queries are not adequately addressed in the context of ranking. This is clearly not appropriate since queries vary largely in semantics [3, 21] and users' search intentions [14, 19]. For example, in a popular taxonomy of Web search, queries can be coarsely categorized as navigational, informational and transactional [4] in terms of search intentions. As another example, queries can vary in terms of relational information needs, including queries for subtopic retrieval [25] and topic distillation [22].

Although query difference is multi-faceted, we observe that query difference usually has tight correlation with the user's different expectation on the result accuracy on different rank positions. Let us elaborate this issue using Broder's "Taxonomy of Web search" [4], which describes query difference based on the search intent of users and classifies queries into three categories: navigational, informational and transactional. In particular, navigational queries are those which are intended to find a specific Web site that the user has in mind; informational searches are intended to find information about a topic; transactional ones are intended to complete some Web-mediated activities. Therefore, for the navigational and transactional query, the user expects high accuracy on the top one retrieved result; while for the informational query the user looks for more relevant documents among top-K rank positions. This kind of position-sensitive query difference requires respective objectives for the ranking model. Specifically, for the navigational and transactional query, the ranking model should aim to rank the exact Web page that the user is looking for on the top position of the result list; while for the informational query, the ranking model should target at presenting a set of Web pages relevant to the topic of the query on the top-K positions of returned results. The above only illustrates one aspect of the issue, we can similarly consider the issue in the context of subtopic retrieval and topic distillation. In particular, for the subtopic retrieval query, the objective of ranking model should be presenting a set of Web pages covering as many subtopics as possible on the top-K positions of result list; while for topic distillation query, the ranking model should focus on ranking a set of Web pages best representing one single topic among top-K rank positions.

In this paper, we propose to incorporate query difference into ranking by introducing query-dependent loss functions in the learning process. Inspired by the diverse ranking objectives implied by various queries, we apply different loss functions to different queries in learning the ranking function. Since it is difficult and expensive in practice to extract individual objective for each query, we make use of query categorization to represent query difference such that each query category stands for one kind of ranking objective. In this paper, we focus on Broder's taxonomy of Web search [4] and develop position-sensitive query dependent loss functions according to this popular query categorization.

Unfortunately, query categorization may or may not be available at learning time. Accordingly, beyond learning the ranking functions with pre-defined query categories, we develop a new method for learning ranking functions jointly with query categorization without prior knowledge on query categorization. In this new method, the ranking function and query categorization use totally disjointed feature sets.

To evaluate the effectiveness of our proposed approach, we derive the position-sensitive query-dependent loss function based on Broder's taxonomy, and apply it to two popular ranking algorithms, RankNet and ListMLE. Experimental analysis is employed to verify that query-dependent loss function can be exploited to boost the accuracy of ranking for Web search. We also make a comparison on the ranking accuracy between the learning method that trains the ranking function with pre-defined query categorization and that learns the ranking function jointly with query categorization. Moreover, we provide analysis and conduct additional experiments to gain deeper understanding on the advantages of ranking with query-dependent loss functions over other query-dependent or query-independent ranking approaches.

The major contributions of this work include:

- Proposing to incorporate query difference into ranking by introducing query-dependent loss functions.
- Introducing a new methods for learning the ranking function jointly with learning query categorization
- Exploiting the position-sensitive query-dependent loss function on a popular query categorization scheme of Web search and applying it to two specific ranking algorithms, RankNet and ListMLE.

The remaining parts of this paper are organized as follows. Section 2 proposes the general idea of incorporating query difference into ranking by introducing query-dependent loss functions. In section 3, we specifically define a position-sensitive query-dependent loss function based on Broder's query categorization for Web search and apply it to two concrete ranking algorithms. Experiments and discussions are presented in section 4. A comparison between our method and other query dependent methods for ranking is discussed in section 5. We conclude the paper and point out future research directions in section 6.

2. INCORPORATING QUERY DIFFERENCE INTO RANKING

In this section, we propose to incorporate query difference into ranking by introducing query-dependent loss function, and we also outline learning methods for ranking with the query-dependent loss functions.

2.1 Query-Dependent Loss Functions

We formalize the problem of building a ranking model as finding a function $f \in \mathcal{F}$, where $\mathcal{F}$ is a given function class, such that f minimizes the risk of ranking in the form of a given loss function L_f. For a general learning to rank approach, the loss function is defined as:

$$L_f = \sum_{q \in Q} L(f), \tag{1}$$

where Q denotes the set of queries in the training data; $L(f)$ denotes a query-level loss function, which is defined on ranking function f and has the same form among all queries.

Inspired by the diverse ranking objectives implied by the queries, we incorporate query difference into the loss function by applying different loss functions to different queries. This kind of query-dependent loss function is defined as:

$$L_f = \sum_{q \in Q} L(f; q), \tag{2}$$

where $L(f;q)$ is the query-level loss function defined on both query q and ranking function f, and each query has its own form of loss function.

However, it is difficult and expensive in practice to define individual objective for each query. Thus, we take advantage of query categorization to represent query difference, which means each query category stands for one kind of ranking objective. In general, we assume there is a query category space, denoted as $\mathbf{C} = \{\mathcal{C}_1, \cdots, \mathcal{C}_m\}$, where $\mathcal{C}_i (i = 1, \cdots, m)$ represents one query category. We also assume a soft query categorization, which means each query can be described as a distribution over this space. We use $P(\mathcal{C}_i|q)$ to denote the probability that query q belongs to the class $\mathcal{C}_i$ with $\sum_{i=1}^{m} P(\mathcal{C}_i|q) = 1$. Thus, the query-dependent loss function of the ranking function f is defined as:

$$\begin{aligned} L_f &= \sum_{q \in Q} L(f; q) && (3) \\ &= \sum_{q \in Q} \left(\sum_{i=1}^{m} P(\mathcal{C}_i|q) L(f; q, \mathcal{C}_i) \right), && (4) \end{aligned}$$

where $L(f;q,\mathcal{C})$ denotes a category-level loss function defined on query q, ranking function f and q's category $\mathcal{C}$.