Expert Systems With Applications 108 (2018) 74–80

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

A learnable search result diversification method

Hai-Tao Zheng*, Jinxin Han, Zhuren Wang, Xi Xiao

Tsinghua-Southampton Web Science Laboratory, Graduate School at Shenzhen, Tsinghua University, China

ARTICLE INFO

Article history:
Received 25 June 2017
Revised 4 April 2018
Accepted 20 April 2018
Available online 24 April 2018

MSC:
00-01
99-00

Keywords:
Explicit search result diversification
Learning model
Markov random fields

ABSTRACT

Search result diversification is to tackle the ambiguous queries and multi-faced information needs. The search result diversification problem can be formalized as a balance between the relevance score and the diversity score. Most previous diversification models utilize a predefined function to calculate the diversity score. The values of parameters need to be tuned by manual experiments. It is time-consuming and hard to reach optimal result in diversity evaluation. Proposing a learnable approach to solve the above problems is a pressing task. Therefore we introduce a Learnable Search Result Diversification model called L-SRD. On this basis, we redefine the diversity function and derive our loss function as the likelihood loss of ground truth generation. Stochastic gradient descent algorithm is employed to optimize the values of parameters. Finally we derive our ranking function to generate the diverse list sequentially. Due to the learning model, the values of parameters are determined automatically and get optimally. The experiments on TREC web tracks show that our approach outperforms several existing diversification models significantly.

1. Introduction

There are many ambiguous queries in search system. The keyword **apple** may refer to the Apple, one of the most famous companies in the world, or the electronics Apple manufactures. It may be the most familiar fruit also. There are many aspects of information needs underlying a simple query. How to produce a good quality diverse result is our main concern.

The existing diversification approaches have been categorized as either implicit approaches or explicit approaches. The implicit approaches assume each document representing its own aspect and promote diversity by selecting documents for different aspects based on the difference of their vocabulary (Carbonell & Goldstein, 1998). It is a less effective model for the reason that it cannot express the inherent meaning well (Agrawal, Gollapudi, Halverson, & Ieong, 2009; Zhai, Cohen, & Lafferty, 2003). The explicit approaches are proposed to overcome the weakness. They explicitly formalize the aspects underlying a query and select documents that cover different aspects. The xQuAD and PM2 are classic explicit models (Dang & Croft, 2012; Santos, Macdonald, & Ounis, 2010a). But they just utilize a predefined function to calculate the diversity score based on query aspects. It is subjective and hard to reach optimal result.

In this paper, we treat the Query Aspect Diversification as a learning problem and propose a Learnable Search Result Diversification (L-SRD) method. We incorporate various features into diversity measurement based on the Markov Random Field (MRF), which enables the integration of various types of features. The values of parameters can be determined automatically, which saves the manual labour, and the parameters are more optimal. Firstly we redefine the diversity function and derive our loss function as the likelihood loss of ground truth generation. Then Stochastic gradient descent algorithm is employed to optimize the values of weights. Finally we derive our ranking function to generate the diverse list sequentially.

We conduct a series of experiments to demonstrate that L-SRD is more effective than other diversification models in terms of the official evaluation metrics including α-nDCG, ERR-IA, NRBP and the classical diversification metrics such as Precision-IA and Aspect Recall (Chapelle, Metlzer, Zhang, & Grinspan, 2009; Clarke et al., 2008; Clarke, Kolla, & Vechtomova, 2009b). Additionally, we get a remarkable performance in robust evaluation.

The main contributions of our work are listed as follows:

1. L-SRD introduces the learning mechanism to the query aspect diversification model. We conduct inference for the loss function based on its sequential selection model, which solves the parameters tuning problem automatically at the same time.

* Corresponding author.
E-mail addresses: zheng.haitao@sz.tsinghua.edu.cn (H.-T. Zheng), hanjx16@mails.tsinghua.edu.cn (J. Han), wang-zr14@mails.tsinghua.edu.cn (Z. Wang), xiaox@sz.tsinghua.edu.cn (X. Xiao).

https://doi.org/10.1016/j.eswa.2018.04.029
0957-4174/

2. We utilize the Markov Random Field to integrate different types of features to address the diversity measurement problem for query aspect search result diversification.
3. We propose a sequential prediction method, which selects the best document from candidate set by maximizing ranking score.
4. We conduct extensive experiments to verify L-SRD achieve better performance comparing with the existing diversification methods.

The remainder of this paper is organized as follows. Section 2 introduces the current research situation on the search result diversification. Section 3 describes the definition of the loss function and the estimation of parameters. Sections 4 and 4.3 detail the experiments setup on the TREC web track and their evaluations. In Section 5, we summarize our achievements and give future works.

2. Related work

Search result diversification has a wide range of applications, such as patent search (Kim & Croft, 2015), legal information retrieval (Koniaris, Anagnostopoulos, & Vassiliou, 2017) and so on. The process of diversification can be characterized as a bidirectional optimization problem, in which one seeks to maximize the overall relevance of a document ranking to multiple query aspects, while minimizing its redundancy (Santos, Macdonald, & Ounis, 2010b). In particular, the existing approaches can be categorized as either implicit or explicit making a difference in how they account for the query aspects (Santos, Peng, Macdonald, & Ounis, 2010c).

The basic assumption of implicit diversification approaches is that dissimilar documents are more likely to satisfy different information needs. The most representative approach in maximal marginal relevance (MMR) method and its probabilistic variants is shown as follows (Zhai et al., 2003):

$$S_{MMR}(q,d,c) = (1-\lambda)S^{rel}(d,q) - \lambda \max_{d_j \in C} S^{div}(d,d_j), \tag{1}$$

where S^{rel} and S^{div} represents document $d's$ relevance to the query q and its similarity to a selected document d_j, respectively. To gain high ranking score, a document should not only be relevant, but also be dissimilar from the selected documents. The special process of MMR proposed by Carbonell and Goldstein (1998) is selecting the document iteratively, and meanwhile, both content-based relevance and diversity relation between current selected document and the previously selected documents are considered. Yu et al. (2017) formulate this as a process of selecting and ranking k exemplar documents and utilize linear programming to solve this problem. In summary, they are all implicit approaches without using aspects to mine the underlying aspects, besides, they are a low effective approaches (Drosou & Pitoura, 2010; Santos et al., 2010a).

Explicit approaches make use of the aspects underlying the query to select documents that cover different aspects as far as possible. The algorithms such as IA-select (Agrawal et al., 2009), xQuAD (Santos et al., 2010a) and RxQuAD (Vargas, Castells, & Vallet, 2012) are proposed to reduce redundancy on the aspect levels. These methods select documents that cover more novel aspects. The PM-1 and PM-2 models pay more attention to maintain the proportionality of aspects (Dang & Croft, 2012). They produce the ranked result according to the proportionality of aspects. Intrinsic diversity products a series of successor queries to figure out the appropriate content to cover (Raman, Bennett, & Collins-Thompson, 2013). Wang, Dou, Sakai, and Wen (2016) and Hu and et al. (2015) think the aspects underlying the query should be hierarchical, and propose some hierarchical measures to find the relationships among aspects. Ullah, Shajalal, Chy, and Aono (2016) mine query subtopic by exploiting the word embedding and short-text similarity measure. To conclude, all existing explicit approaches are unsupervised, and the values of parameters need to be tuned by the experiment repeatedly without intention, causing a time-consuming optimizing problem to find the most suitable parameters.

Some learning approaches are also proposed for search result diversification. For example, Zhu, Lan, Guo, Cheng, and Niu (2014) use structural SVM to learn to identify a document subset with maximum word coverage, but they just learn the maximum word coverage and do not mine the aspects underlying the query. Xia, Xu, Lan, Guo, and Cheng (2015) utilize both positive and negative ranking documents to train a maximal marginal relevance model for ranking. Xia, Xu, Lan, Guo, and Cheng (2016) propose a neural tensor network to learn a nonlinear novelty function to select document. However, different from the existing approaches, we use a learnable process to identify features from documents using Markov Random Field. Besides, we redefine the diversity function and derive our loss function as the likelihood loss of ground truth generation to resolve this bidirectional optimization problem.

3. Learning approach for search result diversification

3.1. Mining aspects underlying the query

The key step for Query Aspects Diversification model is mining the aspects underlying the query. With the help of query aspects, we can generate the diverse ranking list by minimizing the redundancy on the basis of the aspects. We mine the query aspects like (Santos et al., 2010a), issuing the query to the commercial search engine (we use Yahoo) and get back the query suggestion result list as the aspects. Nextly, we can use these aspects as a new query to search the candidate document set D and we can get the relevance score between the aspect q_i and each document d in D, which can be formalized as $P(q_i|d)$.

3.2. Topic diversity model

Traditional topic diversity model is a greedy approximation. It sequentially selects the "local-best" document from the candidate document set (Santos et al., 2010a). The original function is formalized as follows:

$$f(d,\bar{S}) = (1-\lambda)P(d|q) + \lambda \sum_{q_i \in Q} P(q_i|q)P(d|q_i)P(\bar{S}|q_i). \tag{2}$$

where d denotes for the current document to be considered in the sequential process, $\bar{S}$ denotes for the unselected document set (equal to the $D\backslash S$ in Fig. 1), q denotes for the query, λ is a balance parameter for a trade-off between relevance and diversity, q_i denotes for the aspects underlying the query q.

As for Eq. (2), the left part corresponds to the relevance score and the right part corresponds to the diversity score. We look forward to redefine the estimation of diversity score $P(\bar{S}|q_i)$. According to the conditional probabilistic formula, the task can be formalized as follows:

$$P(\bar{S}|q_i) = \frac{P(\bar{S},q_i)}{P(q_i)} \overset{rank}{=} P(\bar{S},q_i) \tag{3}$$

where $P(q_i)$ denotes the occurrence rate of aspects q_i corresponding to query q, which is usually regard to be normalized as $1/n$ (n denotes the number of aspects) (Santos et al., 2010a). Because the values of $P(q_i)$ are equal and do not impact on the result of ranking, we neglect $P(q_i)$.

The main concern is how to define feature function for $P(\bar{S},q_i)$. There are many ways to integrate different features, just like linear regression, logistic regression and some other ways. Under our