Information Processing and Management 54 (2018) 507–528

Contents lists available at ScienceDirect

Information Processing and Management

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

Revisiting the cluster-based paradigm for implicit search result diversification

Hai-Tao Yu[*,a], Adam Jatowt[b], Roi Blanco[c], Hideo Joho[d], Joemon M. Jose[e], Long Chen[e], Fajie Yuan[e]

[a] Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan
[b] Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto, Japan
[c] IRLab, Computer Science Department, University of A Coruña, Spain
[d] Research Center for Knowledge Communities, Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan
[e] School of Computing Science, University of Glasgow, Glasgow, UK

ARTICLE INFO

Keywords:
Cluster-based IR
Implicit SRD
Integer linear programming
Affinity propagation

ABSTRACT

To cope with ambiguous and/or underspecified queries, *search result diversification* (SRD) is a key technique that has attracted a lot of attention. This paper focuses on *implicit SRD*, where the subtopics underlying a query are *unknown*. Many existing methods appeal to the greedy strategy for generating diversified results. A common practice is using a heuristic criterion for making the locally optimal choice at each round. As a result, it is difficult to know whether the failures are caused by the optimization criterion or the setting of parameters. Different from previous studies, we formulate implicit SRD as a process of selecting and ranking k exemplar documents through integer linear programming (ILP). The key idea is that: *for a specific query, we expect to maximize the overall relevance of the k exemplar documents. Meanwhile, we wish to maximize the representativeness of the selected exemplar documents with respect to the non-selected documents. Intuitively, if the selected exemplar documents concisely represent the entire set of documents, the novelty and diversity will naturally arise.* Moreover, we propose two approaches *ILP4ID* (Integer Linear Programming for Implicit SRD) and *AP4ID* (Affinity Propagation for Implicit SRD) for solving the proposed formulation of implicit SRD. In particular, ILP4ID appeals to the strategy of bound-and-branch and is able to obtain the optimal solution. AP4ID being an approximate method transforms the target problem as a maximum-a-posteriori inference problem, and the message passing algorithm is adopted to find the solution. Furthermore, we investigate the differences and connections between the proposed models and prior models by casting them as different variants of the cluster-based paradigm for implicit SRD. To validate the effectiveness and efficiency of the proposed approaches, we conduct a series of experiments on four benchmark TREC diversity collections. The experimental results demonstrate that: (1) The proposed methods, especially *ILP4ID*, can achieve substantially improved performance over the state-of-the-art unsupervised methods for implicit SRD. (2) The *initial runs, the number of input documents, query types, the ways of computing document similarity, the pre-defined cluster number* and *the optimization algorithm* significantly affect the performance of diversification models. Careful examinations of these factors are highly recommended in the development of implicit SRD methods. Based on the in-depth study of different types of methods for implicit SRD, we provide additional insight into the cluster-based paradigm for implicit SRD. In particular, how the methods relying on greedy

* Corresponding author.
E-mail addresses: yuhaitao@slis.tsukuba.ac.jp (H.-T. Yu), adam@dl.kuis.kyoto-u.ac.jp (A. Jatowt), rblanco@udc.es (R. Blanco), hideo@slis.tsukuba.ac.jp (H. Joho), joemon.jose@glasgow.ac.uk (J.M. Jose), Long.Chen@glasgow.ac.uk (L. Chen), f.yuan.1@research.gla.ac.uk (F. Yuan).

https://doi.org/10.1016/j.ipm.2018.03.003
Received 22 August 2017; Received in revised form 25 December 2017; Accepted 7 March 2018

strategies impact the performance of implicit SRD, and how a particular diversification model should be fine-tuned.

1. Introduction

Accurately and efficiently satisfying user information requests by search engines is still far from being a solved problem. A key issue is that users tend to submit short and often ambiguous or underspecified queries; for example, the common query *Lord of the Rings* may refer to the movie series or the book. Furthermore, when it comes to the movies, users may be interested in a variety of possible aspects including the cast, reviews, price of dvds, etc. Correctly determining users' preferences is however difficult. As a remedy, one possible solution is to apply *search result diversification* (SRD) technique, which relies on providing a diversified result set so as to maximize the likelihood that an average user will find documents relevant to her particular search need. Considering the above-mentioned movie example such solution should generate an optimized result list that covers the key possible aspects like *book, movie, dvd*. According to *whether the subtopics (i.e., different information needs) underlying a query are given beforehand or not*, the task of SRD can be distinguished into *implicit SRD* and *explicit SRD*. The distinguishing characteristics of the implicit SRD is that the possible subtopics underlying a query are *unknown*. Noteworthy, finding a group of subtopic strings that covers well all the possible information needs behind the query is a challenging task. In most realistic scenarios explicit subtopics are not available (Kim & Lee, 2015), neither is the training data for supervised methods (e.g., Burges et al., 2005; Cao, Qin, Liu, Tsai, & Li, 2007; Jiang et al., 2017; Radlinski, Kleinberg, & Joachims, 2008; Xia, Xu, Lan, Guo, & Cheng, 2015; Xia, Xu, Lan, Guo, & Cheng, 2016; Xia et al., 2017; Yue & Joachims, 2008). In such scenarios the technique of implicit SRD is then commonly used, instead, for the purpose of diversifying the results and satisfying users' search intents. Consequently, in this paper *we focus on the implicit diversification methods* instead of the explicit SRD or on supervised methods for search result diversification.

The state-of-the-art methods for implicit SRD can be differentiated according to their solutions for the following key problems: (1) how to represent diversity; (2) how to balance the notions of the relevance and diversity, and (3) how to generate the final result list. For example, the well-known Maximal Marginal Relevance (*MMR*) model (Carbonell & Goldstein, 1998) measures the diversity of a document d_i based on the maximum similarity between d_i and the previously selected documents to approach the first challenge and in order to balance relevance and diversity, *most of the existing methods utilize a trade-off parameter* λ. Finally, for generating the desired result list *the common practice is using the reedy strategy that follows a heuristic criterion of making the locally optimal choice at each round* (Carbonell & Goldstein, 1998; Dang & Croft, 2012; Santos, Macdonald, & Ounis, 2010; Zuccon, Azzopardi, Zhang, & Wang, 2012).

Despite the success achieved by the state-of-the-art methods, there are several issues and problems that need further exploration. The key underlying drawback of the state-of-the-art approaches is that the commonly used greedy strategy works well on the premise that the preceding choices are optimal or close to the optimal solution. However, in many cases, this strategy fails to guarantee the optimal solution. A natural question arises then: *to what extent does the greedy solution affect the performance of implicit SRD*? Moreover, when conducting experimental analysis, a single weighting model (e.g., language model with Dirichlet smoothing (Zhai & Lafferty, 2004)) is commonly adopted to perform the initial retrieval of results. Since the initially retrieved documents (e.g., top-m documents) are then further used to test diversification models, different initial runs should have significant impact on the performance of these diversification models. Furthermore, the effects of the key parameters: m (i.e., the number of used documents) and k (i.e., the predefined cluster number) on the performance of a diversification model are crucial and should be explored in details. The same criterion applies to the examination of the effect of the different query types on the quality of results. To the best of our knowledge all these key points have not been sufficiently investigated in most of the previous studies on implicit SRD.

The aforementioned drawbacks motivate us to address the task of the implicit SRD in a novel way. In particular, we propose a concise integer linear programming (ILP) formulation for implicit SRD. Based on such formulation, we introduce two different approaches to find the desired solution. One is an approximate method based on message passing called *AP4ID*. The other is an exact method, called *ILP4ID*, which is based on the strategy of bound-and-branch, under which the exactly optimal solution can be obtained and validated. Finally, we compare the effectiveness of the proposed approaches against the state-of-the-art algorithms using the standard TREC diversity collections. The experimental results prove that both *AP4ID* and *ILP4ID* can improve performance over the baseline methods in terms of the standard diversity metrics.

The main contributions of this paper are as follows:

1. We present a concise ILP formulation for implicit SRD which allows for the exact solution of the objective function (Eq. 12) to be obtained. On the one hand, two different approaches *AP4ID* and *ILP4ID* are proposed to find the desired solution. The proposed method *ILP4ID* can lead to substantially improved performance than the state-of-the-art unsupervised methods. The experimental results also demonstrate how much accuracy has been lost due to the usage of an approximation method (e.g., compared with the method Zuccon et al., 2012). On the other hand, the flexibility of the proposed formulation allows for further extensions by simply altering the constraints.
2. Different from prior studies, we thoroughly investigate the effects of a series of factors on the performance of a diversification model. Our main finding is that some factors, such as *different initial runs, the number of input documents, query types, the ways of computing document similarity* and *the predefined cluster number* greatly affect the effectiveness of diversification models for implicit SRD. Careful examinations of these factors are highly recommended in the development of implicit SRD methods. Based on the