Information Processing and Management 51 (2015) 773–785

Contents lists available at ScienceDirect

Information Processing and Management

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

Subtopic mining using simple patterns and hierarchical structure of subtopic candidates from web documents☆

Se-Jong Kim*, Jong-Hyeok Lee

Department of Computer Science and Engineering, Pohang University of Science and Technology (POSTECH), 77 Cheongam-Ro, Nam-Gu, Pohang 790-784, Republic of Korea

ARTICLE INFO

Article history:
Received 29 October 2014
Revised 2 July 2015
Accepted 6 July 2015
Available online 28 August 2015

Keywords:
Search intention
Subtopic mining
Hierarchical structure

ABSTRACT

The intention gap between users and queries results in ambiguous and broad queries. To solve these problems, subtopic mining has been studied, which returns a ranked list of possible subtopics according to their relevance, popularity, and diversity. This paper proposes a novel method to mine subtopics using simple patterns and a hierarchical structure of subtopic candidates. First, relevant and various phrases are extracted as subtopic candidates using simple patterns based on noun phrases and alternative partial-queries. Second, a hierarchical structure of the subtopic candidates is constructed using sets of relevant documents from a web document collection. Finally, the subtopic candidates are ranked considering a balance between popularity and diversity using this structure. In experiments, our proposed methods outperformed the baselines and even an external resource based method at high-ranked subtopics, which shows that our methods can be effective and useful in various search scenarios like result diversification.

1. Introduction

The advent of smart devices has significantly influenced the web search environment. Unlike in the era of personal computers, enhanced searching services are required to obtain accurate search results because users tend to simplify queries for their own convenience. In fact, many web queries are short and unclear. Some users do not choose appropriate words for a web search, and others omit specific terms needed to clarify search intentions, because it is not easy for users to express their search intentions explicitly through keywords. This intention gap between users and queries results in queries which are ambiguous and broad. In the case of ambiguous queries, users may get results quite different from their intentions; as for broad queries, results may not be as specific as users expect.

As one of the solutions for these problems, web search engines have provided query suggestion services such as autocomplete and related queries. Query suggestion gives new queries to help a user explore and express his information need (search intention) for a given query (Bhatia, Majumdar, & Mitra, 2011; Jones, Rey, Madani, & Greiner, 2006). However, query suggestion does not explicitly consider popularity and diversity of suggested queries. For this reason, as a new solution, subtopic mining is proposed, which can find possible subtopics (suggested queries) for a given query and return a ranked list of them in terms of their relevance, popularity, and diversity (Fig. 1).

☆ A preliminary version of this work was presented in Kim, Shin, and Lee (2013).

* Corresponding author. Tel.: +82 54 279 5656.
E-mail addresses: sejong@postech.ac.kr (S.-J. Kim), jhlee@postech.ac.kr (J.-H. Lee).

http://dx.doi.org/10.1016/j.ipm.2015.07.001
0306-4573/

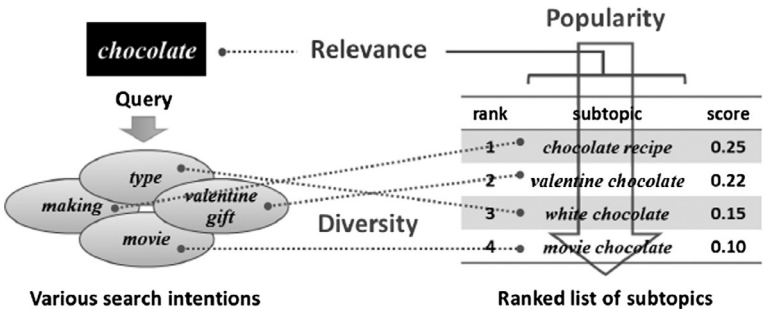

Fig. 1. Relevance, popularity, and diversity in subtopic mining.

According to NTCIR-9 subtopic mining task (Song et al., 2011), a subtopic of a given query is a query that disambiguates and specifies the search intention of the original query. For example, if a query is "*chocolate*," its hyponyms (sub-class words) "*white chocolate*" and "*dark chocolate*" can be a subtopic to specify one search intention "*chocolate type*." In contrast, a hypernym (super-class word) of a given query is excluded from the subtopic candidates because it broadens the search intention. The popularity and diversity consider only the subtopics which are relevant to the query, and are estimated for each of the subtopics and the set of the subtopics, respectively.

Subtopic mining can be used to improve the results of various search scenarios, such as personalized search and search result diversification. Especially, for an explicit approach to search result diversification, subtopic mining can be an important part to find as various subtopics as query aspects that are used to retrieve relevant documents for the query. The study of search result diversification is divided into explicit approach and implicit approach (Dang & Croft, 2012; Dang & Croft, 2013; Santos, Macdonald, & Ounis, 2010). The explicit approach finds explicitly the subtopics (topic-level) or relevant words (term-level) for a given query first, and retrieves relevant documents for each of them. As a result, the set of retrieved documents satisfies the diversity because they contain various contents for the query. The implicit approach retrieves various documents related to a given query without taking the step of finding the subtopics or relevant words for the query. Typically, the explicit approach does not directly focus on the diversity of subtopics or relevant words, whereas it focuses on the diversity of relevant documents considering the novelty (minimum redundancy) (Santos et al., 2010) and proportionality (popularity distribution) (Dang & Croft, 2012). Therefore, subtopic mining could be considered a breakthrough for the improvement of topic-level search result diversification.

Most of previous methods relied on query logs to find subtopics (Baeza-Yates, Hurtado, & Mendoza, 2005; Beeferman & Berger, 2000; Fujita, Uchiyama, Dupret, & Baeza-Yates, 2010; Huang, Chien, & Oyang, 2003; Jones et al., 2006; Ma, Lyu, & King, 2010; Santos et al., 2010; Santos, Macdonald, & Ounis, 2011; Xue et al., 2011; Zhang, Lu, & Wang, 2011; Zhu, Guo, Cheng, Du, & Shen, 2011). The query logs are particularly useful resources for existing search engines to build bipartite graphs of ⟨query, clicked URL⟩, and thus to produce suggested queries. However, most of them are not available to external researchers, and data sparseness occurs because rare queries may be few or non-existent in query logs. Meanwhile, the previous methods focused more on the relevance to rank subtopics than the popularity (Fujita et al., 2010; Huang et al., 2003; Jones et al., 2006; Ma et al., 2010; Zhu et al., 2011), or made no distinction between them (Baeza-Yates et al., 2005; Santos et al., 2011; Xue et al., 2011). Though a useful factor for document ranking, the relevance does not reflect the popularity of subtopics because the relevance only checks whether a subtopic is similar or related to a given query. For example, if a query is "*typhoon*," "*eye of a typhoon*" is the higher related subtopic than "*typhoon damage*." However, general users get more interested in "*typhoon damage*" than "*eye of a typhoon*." In other words, the relevance is the preceding condition to find subtopics, and afterwards the popularity is considered to rank them. Furthermore, as an essential issue, the popularity and the diversity are not proportional to each other. For example, if some subtopics belong to only a few search intentions with high popularity, their diversity is low. Good subtopics must be relevant to a given query and satisfy both the high popularity and high diversity.

To solve these issues, we propose a novel method to mine subtopics using simple patterns and a hierarchical structure of subtopic candidates based on relevant documents. Our contributions are as follows:

- Our method uses only web document collection instead of query logs and external resources.
- To find various relevant subtopic candidates, we extract as many "*understandable*" phrases as possible from the web documents, which fully or partially match the original query through simple patterns. These patterns are based on noun phrases and alternative partial-queries to the original query, and this approach is easily adaptable to other languages.
- Our work proposes a hierarchical structure of subtopics to maintain a balance between popularity and diversity. This structure has relatively small groups of subtopics, which cover a wide variety of search intentions of the query and also cover their more specific ones to improve the diversity. There is also an inheritance of popularity between subtopics. Using this structure, we get a ranked list of subtopics satisfying both the high popularity and high diversity.