Query-Oriented Keyphrase Extraction

Minghui Qiu, Yaliang Li, and Jing Jiang

School of Information Systems
Singapore Management University
{minghui.qiu.2010,ylli,jingjiang}@smu.edu.sg

Abstract. People often issue informational queries to search engines to find out more about some entities or events. While a Wikipedia-like summary would be an ideal answer to such queries, not all queries have a corresponding Wikipedia entry. In this work we propose to study query-oriented keyphrase extraction, which can be used to assist search results summarization. We propose a general method for keyphrase extraction for our task, where we consider both phraseness and informativeness. We discuss three criteria for phraseness and four ways to compute informativeness scores. Using a large Wikipedia corpus and 40 queries, our empirical evaluation shows that using a named entity-based phraseness criterion and a language model-based informativeness score gives the best performance on our task. This method also outperforms two state-of-the-art baseline methods.

Keywords: Keyphrase extraction, phraseness, informativeness, language model.

1 Introduction

Online searches generally fall into three categories, namely, informational, navigational and transactional [5]. A recent study has found that the majority of online queries are informational [14], which intend to locate information pertaining to a certain topic. A typical type of informational queries is to simply find out more about a topic such as an entity or an event. For this type of informational queries, instead of showing a ranked list of URLs in the traditional way, a short summary article for the query topic might be a better form to present the search results, from which users can easily digest and further explore. Indeed, recently Google's search results started to include a summary page on the right hand side backed by Google's *Knowledge Graph*, demonstrating the need to automatically summarize information related to a query. But it still remains a challenging task to automatically generate open-domain summaries without supervision.

In this paper, we take a less ambitious step and propose to study the task of finding related keyphrases given a query. This kind of query-oriented keyphrases can be useful for search in a number of ways. For example, to generate an extractive summary of the search results, one may select sentences that maximize the coverage of these related keyphrases. Related keyphrases can also serve as anchor points for further navigation from the original search results in exploratory search.

In Table 1 we show a sample output of our proposed task for the query *Pixar*, where the top-10 keyphrases returned by the best configuration of our proposed method are listed. We can see that these top keyphrases are highly relevant to the query.

Y. Hou et al. (Eds.): AIRS 2012, LNCS 7675, pp. 64–75, 2012.

Table 1. Top-10 keyphrases for the query "Pixar" returned by our method. The descriptions are given by the authors.

Returned Keyphrase	Description
pixar animation studios	full name of the company
john lasseter	CCO of Pixar and Walt Disney animation studios
walt disney pictures	parent company of Pixar
bob iger	CEO of the Walt Disney Company
pixar story	-
walt disney company	owner of Pixar
andrew stanton	director of some Pixar movies
brad bird	director of some Pixar movies
luxo jr.	a Pixar movie
tow mater	the deuteragonist in the Pixar movie "cars"

To find related keyphrases, we transform the task into a query-oriented keyphrase extraction problem where the goal is to extract keyphrases from a set of documents relevant to the given query. Keyphrase extraction has been extensively studied before [12,20,19,22,21,13]. Existing work includes both supervised and unsupervised approaches. Because of the nature of our task, an unsupervised keyphrase extraction method is needed.

Following a framework proposed by Tomokiyo and Hurst [19], we propose a general method for our task which considers both *phraseness* and *informativeness* of a candidate keyphrase. We consider three phraseness criteria based on language models, noun phrase chunking and named entity recognition, respectively. We also consider four informativeness scores using phrase-level Tf-Idf, sum of word-level Tf-Idf, average of word-level Tf-Idf and language models, respectively.

We evaluate the various combinations of the criteria and compare our method with state-of-the-art baselines using 40 queries and a large Wikipedia corpus. We use ground truth both annotated by humans and automatically obtained from Wikipedia articles for evaluation. Experimental results show that the named entity-based phraseness criterion is the best for our task, and the language model-based informativeness score gives the best keyphrase ranking. This configuration of our method outperforms the two state-of-the-art baseline methods we consider.

Our main contributions are twofold. First, we propose to study a new task of query-oriented keyphrase extraction and provide a general solution based on phraseness and informativeness. Second, we empirically compare different phrasesness and informativeness criteria for this task and find a solution better than the baselines that represent the state of the art of keyphrase extraction.

2 Related Work

To the best of our knowledge the task of query-oriented keyphrase extraction has not been well studied. A related line of work is search results clustering, where oftentimes labels for clusters of documents are automatically generated [15,23]. These labels can be seen as phrases related to the query. A major difference between this line of work and our task is that our related keyphrases are not meant as topical labels for a cluster of documents but rather important concepts related to the query. For example, given