A Method of Subtopic Classification of Search Engine Suggests by Integrating a Topic Model and Word Embeddings

Tian Nie, Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan
Yi Ding, Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan
Chen Zhao, Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan
Youchao Lin, Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan
Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan

ABSTRACT

The background of this article is the issue of how to overview the knowledge of a given query keyword. Especially, the authors focus on concerns of those who search for web pages with a given query keyword. The Web search information needs of a given query keyword is collected through search engine suggests. Given a query keyword, the authors collect up to around 1,000 suggests, while many of them are redundant. They classify redundant search engine suggests based on a topic model. However, one limitation of the topic model based classification of search engine suggests is that the granularity of the topics, i.e., the clusters of search engine suggests, is too coarse. In order to overcome the problem of the coarse-grained classification of search engine suggests, this article further applies the word embedding technique to the webpages used during the training of the topic model, in addition to the text data of the whole Japanese version of Wikipedia. Then, the authors examine the word embedding based similarity between search engines suggests and further classify search engine suggests within a single topic into finer-grained subtopics based on the similarity of word embeddings. Evaluation results prove that the proposed approach performs well in the task of subtopic classification of search engine suggests.

KEYWORDS

Clustering, Overview, Search Engine Suggest, Topic Model, Word Embeddings

INTRODUCTION

This paper addresses the issue of how to overview the knowledge of a given query keyword. We especially focus on concerns of those who search for Web pages with a given query keyword. The approach we take in this paper is to collect Web search information needs of a given query keyword through search engine suggests. Fig. 1 shows an example of presenting search engines suggests given a query keyword “shu-katsu” (job hunting). Here, the search engine collected user search logs including the query keyword “shu-katsu” (job hunting) and then presents suggests keywords such as...
“meeru” (e-mail), “mensetsu” (interview), “2-chan” (2channel), and “meiku” (makeup) which have strong relation to the query term “shu-katsu” (job hunting). After collecting those suggest keywords from a search engine, this paper studies how to efficiently overview the whole list of Web search information needs of a given query keyword. In the case of a Japanese search engine, we collect up to around 1,000 suggests given a query keyword. Some of those 1,000 suggests are quite redundant in that they originate from almost the same Web search information needs. In order to aggregate such redundant search engine suggests, we first take an approach of classifying search engine suggests based on a topic model (Blei, Ng, & Jordan, 2003). More specifically, we collect Web pages by specifying the query keyword and each search engine suggest into a set of Web pages. Then, we apply a topic model to the set of collected Web pages and obtain a certain number of topics. Next, by assigning a topic to each Web page, we collect the set of Web pages that are assigned to each topic. Also, for each topic, we collect search engine suggests that are specified when retrieving those Web pages assigned to the topic into the set of suggests. Here, we consider the set of collected search engine suggests for each topic as clusters of search engine suggests and thus consider them as the result of classifying search engine suggests by modeling topics of Webpages collected with suggests. Fig. 2 shows the overview of the procedure so far of classifying search engine suggests based on a topic model. Redundant search engine suggests are successfully grouped together into a cluster by a topic model. However, one limitation of the topic model based classifying of search engine suggests is that the granularity of the topics, i.e., the classification of search engine suggests, is too coarse. In order to overcome the problem of the coarse-grained classification of search engine suggests, this paper further applies the word embedding technique (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) to the Web pages used during the training of the topic model, in addition to the text data of the whole Japanese version of Wikipedia. Then, we examine the word embedding based similarity between search engines suggests and further classify search engine suggests within a single topic into finer-grained subtopics based on the similarity of word embeddings. Evaluation results prove that the proposed approach performs well in the task of subtopic classifying of search engine suggests.

COLLECTING SEARCH ENGINE SUGGESTS

For a given query keyword, we specify about 100 types of Japanese hiragana characters to Google Search and then collect at most about 1,000 suggests. About 100 types of Japanese hiragana characters include Japanese alphabet consisting of about 50 characters, voiced and semi-voiced variants of voiceless characters, and Youon. For example, when we type in “shu-katsu a” (“job hunting”, “a”) into the Web search window, we can collect suggests which start with the reading character “a” such
10 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:


[www.igi-global.com/e-resources/library-recommendation/?id=163](www.igi-global.com/e-resources/library-recommendation/?id=163)

Related Content

Stock Price Prediction Using Fuzzy Time-Series Population Based Gravity Search Algorithm

Power Transmission Analysis in Wireless Sensor Networks Using Data Aggregation Techniques
www.igi-global.com/chapter/integrated-product-life-cycle-management/29433?camid=4v1a

A Formal Approach for Securing XML Document
www.igi-global.com/article/formal-approach-securing-xml-document/39008?camid=4v1a