Chapter 9
An Automatic Machine Learning Method for the Study of Keyword Suggestion

Lin-Chih Chen
National Dong Hwa University, Taiwan

ABSTRACT

Keyword suggestion is an automatic machine learning method to suggest relevant keywords to users in order to help users better specify their information needs. In this chapter, the authors adopt two semantic analysis models to build a keyword suggestion system. The suggested keywords returned from the system not only with a certain semantic relationship, but also with a similarity measure. The benefit of the authors' method is to overcome the problems of synonymy and polysemy over the information retrieval field by using a vector space model. This chapter shows that using multiple semantic analysis techniques to generate relevant keywords can give significant performance gains.

INTRODUCTION

Meeting users’ search requirement is always one of the most fundamental and challenging issues in the design of search engines. What makes this issue challenging is that most Internet users always give only short queries. According to the analysis of search engine transaction logs, the average length of queries is about 2.3 words (Silverstein, Henzinger, Marais, & Moricz, 1998; Spink, Wolfram, Jansen, & Saracevic, 2001). Thus, it is not simple to find out real search goal from such short queries. In order to effectively deal with the problem of short query, incorporating some kind of the keyword suggestion mechanisms (Belkin, 2000) has become a commonly practice in the search engine design (Google, 2006; Microsoft, 2006; Yahoo, 2006).

Keyword suggestion is a kind of Information Retrieval (IR) technique that attempts to suggest relevant keywords to help the users formulate more effective queries and reduce unnecessary search steps. According to related research as seen in...
An Automatic Machine Learning Method for the Study of Keyword Suggestion

(Abhishek & Hosanagar, 2007a; Yifan Chen, Xue, & Yu, 2008; Ferragina & Guli, 2008; Janruang & Kreesuradej, 2006; Joshi & Motwani, 2006; Wang, Mo, Huang, Wen, & He, 2008), this technique can be broadly classified into three categories: log analysis, proximity analysis, and snippet analysis. The category of log analysis analyzes the content of query logs to suggest relevant keywords (Abhishek & Hosanagar, 2007a; Bartz, Murthi, & Sebastian, 2006; Google, 2006; Lee, Huang, & Hung, 2007; Mei, Zhou, & Church, 2008; Yahoo, 2006). The category of proximity search sends the seed keyword to several search engines and expands new suggested keywords in its proximity range (Abhishek & Hosanagar, 2007b; Yifan Chen, et al., 2008; Joshi & Motwani, 2006). The category of snippet analysis first collects the snippets that are summarized by remote search engines; it then uses several snippet cleaning and pattern matching techniques to extract relevant keywords (Ferragina & Guli, 2008; Janruang & Kreesuradej, 2006; Wang, et al., 2008).

Two additional problems with most traditional keyword suggestion methods are the low coverage and the lack of disambiguation ability. In some cases, two relevant keywords never occur with each other. They will not be found by the traditional methods. In other cases, a keyword may have more than one meaning. A very famous example is that “apple” has at least two meanings: it can be either fruit or corporation. The relevant keywords of apple for these two meanings are obviously different. The traditional methods cannot distinguish between these two meanings and the suggested keywords may be a mixture of both meanings (Yifan Chen, et al., 2008).

The main purpose of our system is not only to suggest relevant and important keywords, but also to measure the degree of similarity between keywords. The screen dump of our system is shown in Figure 1. Our system is based on two semantic analysis models, including Latent Semantic Indexing (LSI) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) and Probabilistic LSI (PLSI) (Hofmann, 1999). The bases of these two semantic analysis models are based on the concept of automatic machine learning to

Figure 1. The screen dump of the authors’ system in response to the search query is “mobile phone” (accessed from http://cayley.sytes.net/li_new)
Related Content

The Formal Design Model of a File Management System (FMS)
www.igi-global.com/article/formal-design-model-file-management/53165?camid=4v1a

A Novel Convolutional Neural Network Based Localization System for Monocular Images
Chen Sun, Chunping Li and Yan Zhu (2019). International Journal of Software Science and Computational Intelligence (pp. 38-50).
www.igi-global.com/article/a-novel-convolutional-neural-network-based-localization-system-for-monocular-images/233522?camid=4v1a

Cooperative Encoding Strategy for Gate Array Placement
Hong-Bo Wang, Qing-Dong Su and Ruolei Zeng (2018). International Journal of Software Science and Computational Intelligence (pp. 29-43).
www.igi-global.com/article/cooperative-encoding-strategy-for-gate-array-placement/223493?camid=4v1a

Discovering Attribute-Specific Features From Online Reviews: What Is the Gap Between Automated Tools and Human Cognition?
www.igi-global.com/article/discovering-attribute-specific-features-from-online-reviews/202951?camid=4v1a