Expert Systems with Applications 41 (2014) 663–670

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Finding keywords in blogs: Efficient keyword extraction in blog mining via user behaviors

Yi-Hui Chen [a], Eric Jui-Lin Lu [b,*], Meng Fang Tsai [b]

[a] Department of Applied Informatics and Multimedia, Asia University, Taichung 41354, Taiwan, ROC
[b] Department of Management Information Systems, National Chung Hsing University, Taichung 40227, Taiwan, ROC

ARTICLE INFO

Keywords:
Blog mining
User intention
Co-keyword
Blog Connect
Full-text keyword retrieval procedure

ABSTRACT

Readers are becoming accustomed to obtaining useful and reliable information from bloggers. To make access to the vastly increasing resource of blogs more effective, clustering is useful. Results of the literature review suggest that using linking information, keywords, or tags/categories to calculate similarity is critical for clustering. Keywords are commonly retrieved from the full text, which can be a time-consuming task if multiple articles must be processed. For tags/categories, there is also a problem of ambiguity; that is, different bloggers may define tags/categories of identical content differently. Keywords are important not only to reflect the theme of an article through blog readers' perspectives but also to accurately match users' intentions. In this paper, a tracing code is embedded in Blog Connect, a newly developed platform, to collect the keywords queried by readers and then select candidate keywords as co-keywords. The experiments show positive data to confirm that co-keywords can act as a quick path to an article. In addition, co-keyword generation can reduce the complexity and redundancy of full-text keyword retrieval procedures and satisfy blog readers' intentions.

1. Introduction

As a result of easy access to the Internet, Web users can readily share their experiences and express their ideas. Because it is convenient for Web users to post content on blogs, blogs have become a major platform through which to share information. As pointed out in 2011 Blogging Statistics (Treanor, 2011), the number of blog sites increased from 3 million in 2004 to 164 million in July 2011. The importance of blogs was revealed in the Technorati's 2010 report (Sobel, 2010), which showed that 40 percent of blog readers trust bloggers' opinions more than they trust the mainstream media. Also, 48 percent of bloggers believe that net surfers will receive more of their ideas, news, and entertainment from blogs in the next five years than from traditional media. Although bloggers and blog readers have mainly visited their friends' blogs in the past, Technorati's 2011 report (Blogosphere, 2011) pointed out that this is no longer the case. Instead, more and more bloggers share information content. Some potential business activities can be mined by blogs (Chen, Tsai, & Chan, 2008). Despite the importance of blogs in information sharing, each blog is still considered an isolated island (Bojārs, Breslin, Peristeras, Tummarello, & Decker, 2008). That is, no connection or relationship between any two blogs is assumed, unless they have been manually created using Blogrolls, citation links, or comments (Gao & Lai, 2010; Lu & Zhu, 2010). If all related blogs could somehow be "auto-magically" connected, it might result in a breakthrough in information sharing.

To create relationships among blogs, the general practice is to first identify a set of keywords for each blog post and then to use these keyword sets to calculate the similarity between any two blog posts (Hu & Bin, 2006; Larsen & Aone, 1999; Markines et al., 2009; Ohtsuki, Matsutoka, Matsunaga, & Furui, 1998). Similar blog posts can be classified/categorized into the same group. In general, there are two approaches to identifying a keyword set for each blog: One is to extract keywords from the blog content (Elsas, Arguello, Callan, & Carbonell, 2008; Gao & Lai, 2010; Kuzar & Navrat, 2010; Qamra, Tseng, & Chang, 2006; Singh & Joshi, 2011) and the other is to retrieve tags or categories defined by bloggers (Hope, Wang, & Barkataki, 2007; Lu & Lee, 2008; Markines et al., 2009; Srinivas, Tandon, & Varma, 2010; Tsai, 2011; Zhang et al., 2009). To extract keywords from the blog content, a full-text keyword retrieval process (FKRP) is generally required. The FKRP typically consists of the following steps: Download the *i*th article, denoted as A_i, from a blog site, scan the full-length article, obtain the article without HTML tags, depicted as $\bar{A}_i$, and get the main content of $\bar{A}_i$ by removing irrelevant data, denoted as $\tilde{A}_i$. Finally, the FKRP tokenizes or segments the contents of $\bar{A}_i$ and $\tilde{A}_i$, and then selects the candidate keywords for $\bar{A}_i$ and $\tilde{A}_i$, depicted as $FKRP(\bar{A}_i)$ and $FKRP(\tilde{A}_i)$, respectively. However, the FKRP process is complicated and time-consuming. Furthermore, the contents of blog

* Corresponding author. Tel.: +886 4 22840864x696; fax: +886 4 22857173.
E-mail addresses: chenyh@asia.edu.tw (Y.-H. Chen), jllu@nchu.edu.tw (E.J.-L. Lu).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.091

articles may change, which means that the process has to be repeated multiple times.

Because bloggers define tags or categories for blog articles, some research has treated tags/categories as the keywords for these articles. However, since bloggers select and define tags/categories, there is inherent ambiguity in that the same content may be labeled with different tags/categories based on individual bloggers' points of view (Hope et al., 2007; Srinivas et al., 2010).

It is now common practice for users to enter keywords on search engines to find what they want. The "wisdom of crowds" concept (Abhishek & Hosanagar, 2007; Agarwal, Galan, Liu, & Subramanya, 2010; Fuxman, Tsaparas, Achan, & Agrawal, 2008) suggests that queried keywords can be used to represent users' intentions (Jansen, Booth, & Spink, 2007) so it is assumed that some of the candidate queried keywords for a blog article (or co-keywords) can be used to represent the topics of the blog article. To verify this assumption, we developed a platform called Blog Connect (BC). When users visit blog article A_i, through search engines, the tracing code embedded in the BC widget (Lu, 2010) collects users' queried keywords. Suppose that the collected queried keywords for A_i are denoted as QK_i, and the set of co-keywords for A_i, denoted as CK_i, is selected according to the term frequency (TF) or term frequency-inverse documents frequency (TFIDF) (Salton & McGill, 1986) values of each keyword in QK_i. If the keyword set generated by FKRP can be used instead of CK_i, the complicated procedure of FKRP can be simplified. To verify the assumption, the similarity between CK_i and the keyword set of $FKRP(\overline{A}_i)$ and that between CK_i and $FKRP(\widetilde{A}_i)$ are calculated to compare the results regarding whether keywords generated by CK_i are suitable to be used instead of those extracted by FKRP.

Additionally, because blog articles are in the open domain, it is necessary to categorize articles into the specific domain manually in advance while using TFIDF. It is a time-consuming process to classify an article into specific topics. Compared to TF' IDF, TF' does not need to classify articles in advance. Therefore, it is a benefit that no time is required to classify articles into several specific domains manually if the results of TF' can be used to replace those of TF'IDF for blog articles. Also, the accuracy of co-keywords selected by TF and TFIDF are compared in the experiments.

In this paper, four evaluators were used to compute the similarity among CK_i, $FKRP(\overline{A}_i)$, and $FKRP(\widetilde{A}_i)$, namely, projection overlap (*po*), projection mutual information (*pm*) (Markines et al., 2009), distributional mutual information (*dm*) (Markines et al., 2009), and Kendall's tau coefficient (Kendall, 1938). To ensure that blog articles were valid, we considered the users' rauding time (Carver, 1997) to filter the articles in the experiments; rauding time refers to a blog reader who visits a blog article and then leaves immediately.

This paper is organized as follows: Related works are briefly described in Section 2. In Section 3, the proposed scheme is presented, including the architecture of Blog Connect, the dataset collected by the BC widget, and the FKRP process. The details of framework similarities are discussed in Section 4. The experimental results and analysis are shown in Section 5. Finally, conclusions and directions for future work are offered in Section 6.

2. Literature review

The blog is one of the main platforms for information sharing. Blog clustering is widely employed to help users efficiently search and acquire blog information (Salton & McGill, 1986). Blog clustering calculates the similarity between blog articles, and similarity is mainly calculated based on articles' linking information (Kim, Candan, & Tatemura, 2007; Lu & Zhu, 2010), keywords (Gao & Lai, 2008; Juffinger & Lex, 2009; Kuzar & Navrat, 2010; Qamra et al., 2006; Singh & Joshi, 2011), or tags/categories (Markines et al., 2009; Srinivas et al., 2010). The linking information is contained within a blog article, such as blog links, citation links, or comments. Lu and Zhu (Lu & Zhu, 2010) claimed and proved that blog links can be viewed as a blogger's browsing behavior, which reflects the user's interest to a certain extent. Based on their browsing behavior, one can discover bloggers' preferences.

The simple way to achieve blog clustering is to compare the similarities of any two blogs according to topics, keywords, and content of articles (Gao & Lai, 2008; Juffinger & Lex, 2009; Kuzar & Navrat, 2010; Qamra et al., 2006; Singh & Joshi, 2011). Because the topic is always too short to completely represent an article and no keywords are defined in the blog article, FKRP is a general approach to selecting the terms as keywords to represent a quick path to a blog article. In the FKRP approach, the entire content of the blog article is first scanned to pick out terms and then the term frequency of each term as the weight of each term using the TF or TFIDF approach is calculated. Finally, the terms are chosen as co-keywords if the calculated weight is sufficiently high.

FKRP is a precise but a time-consuming method. It is doubtful that FKRP can effectively satisfy the requirements of the vast and ever-increasing resource of blogs. Instead of FKRP, some researchers (Agarwal et al., 2010; Lai, Rajashekar, & Rand, 2011; Markines et al., 2009; Srinivas et al., 2010) have suggested that the tags or categories of blog articles can be treated as keywords. To prove that the tags can be used as keywords, Lai et al. (Lai et al., 2011) applied Kendall's tau coefficient to measure the similarity between the tags and keywords extracted by FKRP, where the tags were defined by users in Amazon and the keywords were extracted from FKRP in Twitter. In addition to Kendall's tau coefficient, Markines et al. (Markines et al., 2009) utilized the three-dimensional folksonomy (i.e., tag, resource, user) to calculate the similarity between any two resources. In Markines et al.'s scheme (Markines et al., 2009), four aggregation methods, including projection, distributional, macro-aggregation, and collaborative, were provided to aggregate the three dimensional folksonomy into four data pools. For each data pool, Markines et al.'s scheme applied six similarity measurements, namely, matching, overlap, Jaccard, dice, cosine, and mutual information, to calculate the similarity between any two datasets. In Markines et al.'s experiments, mutual information provided the most accurate of similarity measurements (Markines et al., 2009). However, the user dimension of folksonomy in Markines et al. (2009) was somewhat diminished after the aggregation. To improve Markines et al.'s scheme, Srinivas et al. (Srinivas et al., 2010) proposed a weighty model to incorporate with the user dimension of folksonomy. Although tags or categories can be used to represent an article, as pointed out in Hope et al. (2007) and Srinivas et al. (2010), this method still suffers from a few problems, such as inherent ambiguity, synonym problems, and homonym problems. To alleviate these problems, Agarwal et al. (Agarwal et al., 2010) provided a blog dictionary including the metadata of tags (or categories) for bloggers to help them select appropriate tags (or categories) for their blog articles.

3. Data description and processing

The BC (Lu, 2010) is a cross-platform system developed to help bloggers analyze incoming flows and create relationships among blogs with similar interests or topics. Registered bloggers can easily obtain a piece of tracing code, called a BC widget, which is written in Javascript and can be embedded in a blog. When blog readers enter query keywords on any search engine and then visit a blog article, the queried keywords, stay time, visit timestamp, and incoming URLs, among other data items, are collected by the BC widget and stored in a BC database.