Journal of Network and Computer Applications 43 (2014) 42–55

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Generating temporal semantic context of concepts using web search engines

Zheng Xu [a,b,*], Yunhuai Liu [a], Lin Mei [a], Chuanping Hu [a], Lan Chen [a]

[a] The Third Research Institute of the Ministry of Public Security, 339 Bisheng Road, Shanghai 201142, China
[b] Tsinghua University, Beijing, China

ARTICLE INFO

Article history:
Received 25 February 2013
Received in revised form
5 March 2014
Accepted 4 April 2014
Available online 5 May 2014

Keywords:
Temporal semantic context
Semantic annotation
Content analysis
Web mining

ABSTRACT

In this paper, the problem of generating temporal semantic context for concepts is studied. The goal of the proposed problem is to annotate a concept with temporal, concise, and structured information, which can reflect the explicit and faceted meanings of the concept. The temporal semantic context can help users learn and understand unfamiliar or newly emerged concepts. The proposed temporal semantic context structure integrates the features from dictionary, Wikipedia, and LinkedIn web sites. A general method to generate temporal semantic context of a concept by constructing its associated words, associated concepts, context sentences, context graph, and context communities is proposed. Empirical experiments on three different datasets including Q–A dataset, LinkedIn dataset, and Wikipedia dataset show that the proposed algorithm is effective and accurate. Different from manually generated context repositories such as LinkedIn and Wikipedia, the proposed method can automatically generate the context and does not need any prior knowledge such as ontology or a hierarchical knowledge base. The proposed method is used on some applications such as trend analysis, faceted exploration, and query suggestion. These applications prove the effectiveness of the proposed temporal semantic context problem in many web mining tasks.

1. Introduction

With the high speed development of the internet, search has emerged as a key technology to facilitate access to information for users. Millions of users submit millions of queries to web search engines such as Google[1] and Yahoo[2]. Web search engines allow users to browse on the web, find related information, or as a starting point for entertainment.

Given a new concept to the user, she/he may use the web search engines to index the web pages, which may help users to learn the concept conveniently. With sophisticated algorithms, web search engines have made accessing information easy. Some researchers (Sparrow et al., 2011) suggest that when faced with difficult questions, people are primed to think about using computers. When people expect to have future access to information, they have lower rates of recall of the information itself and enhanced recall instead for where to access it. In other words, when faced with a new concept, users prefer to use search engines rather than learn it through their own prior knowledge.

Though web search engines have become a major intermediary for seeking information, finding relevant information satisfying a user's needs based on the user's initial search queries has become an increasingly difficult task (Leung et al., 2003), which makes users.

(1) **Make costly efforts to find useful information.** In White and Drucker (2007), the authors give some statistics: about 29% of users will modify their original queries in a search task session; the average number of re-visit web pages in a search task session is 5; about 21% of operations in a search task session are backward (users revisit the web pages).
(2) **Face high cognitive burden.** The average steps of a search task session are 17.7 (White and Drucker, 2007), which means that the users browse 17.7 web pages before finishing the task.
(3) **Become lost in the search task.** The average branches of a search task session are 4.1 (White and Drucker, 2007). This is the number of times a user revisited a previous page on the trail and then proceeds forward to view another page.

Modified original queries, so many backward operations, too many browsed web pages, and inappropriate branches – such obstacles make it difficult to get the relevant and correct information. In our view, the following causes contribute to difficulties in a search task session:

* Corresponding author at: The Third Research Institute of the Ministry of Public Security, 339 Bisheng Road, Shanghai 201142, China.
E-mail address: xuzheng@shu.edu.cn (Z. Xu).
[1] www.google.com.
[2] www.yahoo.com.

http://dx.doi.org/10.1016/j.jnca.2014.04.002
1084-8045/

(1) **Caused by users.** A study by Jansen et al. (1998) proposes that the average length of a query submitted to popular web search engines is only 2.35 terms. The short queries may not be able to describe users' real information needs. Thus, the short queries given by users can hardly bring good search results. The reasons for users submitting short queries are the low length queries lessen user's cognitive burden; users face an open-ended search task; and users have difficulty in formatting proper queries (Pandit and Olston, 2007).
(2) **Caused by search engines.** Web search engines generally adopt a "one-size-fits-all" approach for search results presentation, which does not consider the personal need of the users. Different users need different information from the same queries.
(3) **Caused by concepts.** The short queries submitted by users are usually represented by ambiguous concepts. For example, the user may submit "apple" when she/he wants to buy a product produced by Apple Computer Company. But the concept usually has different meanings. For example, the concept "apple" can be fruit or a computer company. Besides the diverse semantics of the concept, new meanings of the old concept and newly emerged concepts often lessen the accuracy of the search results.

In order to provide an accurate annotation for a concept, the problem of automatically generating temporal semantic context (TSC) for concepts is studied. Explicit and concise information for a concept is provided, which indicates the semantics and hidden meanings of the concept. What is a good semantic context of a concept? Let's see two examples from the Cambridge Advanced Learner's Dictionary and Wikipedia,[3] which are shown in Fig. 1. In Fig. 1, the semantic context of a concept is structured as follows. First, the definition of the concept is presented. Second, some example sentences are given to show the usage of the concept. In addition, a visual thesaurus graph of the concept is given to show some related concepts. Wikipedia provides the disambiguated meaning of the concept besides the definition of the concept. For example, Wikipedia gives the link to the Apple Inc. which is another meaning of the concept "apple".

Analogously, if the structured and semantic related information of a concept are provided, it will be very helpful for her/him to understand and further explore it. Of course, when a concept is temporally changed, the new meaning may add to the concept. For example, "Gangnam" is a region in Seoul, Korea. But with the popularity of the song "Gangnam Style" recently, the concept "Gangnam" may be related to the popular music. Thus, the temporal feature to the semantic context must be added. So, what is a good temporal semantic context? Let's see another example from LinkedIn,[4] which is shown in Fig. 2. In Fig. 2, the experience of "Barack Obama" is listed by time sequence. In different time intervals, the concept "Barack Obama" has different semantic context. Thus, inspired by the annotations from the dictionary, Wikipedia, and LinkedIn, the temporal semantic context should include:

(1) **Example sentences**. Given a concept, the example sentence can help the users understand the context of the concept. Moreover, the example sentences can help users apply the concepts in a real context. This factor can be found from the dictionary.
(2) **Diverse meanings.** Given a concept, the different meanings should be given, which can help users learn and explore the concept. This factor can be found from Wikipedia.
(3) **Semantically related concepts.** Similar to the synonyms or thesauri in a dictionary, related concepts should be added to the temporal semantic contexts.
(4) **Temporal annotations.** In different time intervals, the concept may have different meanings. The appropriate semantic context in different time intervals should be mined. This factor can be found from LinkedIn.

To the best of our knowledge, the temporal semantic context of concepts has not been well addressed in existing work. The detailed analysis of the existing work will be given in the next section. The major contributions of this paper are as follows.

(1) In this work, the problem of generating temporal semantic context of concepts is proposed. A general method to automatically generate structured temporal contexts of a concept including semantically related words, example sentences, diverse meanings, and temporal annotations is given. The proposed TSC structure integrates the features from dictionary, Wikipedia, and LinkedIn web sites, which is helpful for users to understand and explore the concept.
(2) Empirical experiments on three different datasets including Q–A dataset, LinkedIn dataset, and Wikipedia dataset show that the proposed algorithm is effective and accurate. Different from manually generated context repositories such as LinkedIn and Wikipedia, the proposed method can automatically generate the context. Moreover, the proposed method does not need any prior knowledge such as ontology or a hierarchical knowledge base such as WordNet.[5]
(3) Some applications using the proposed TSC method are given. The proposed method can be used on trend analysis, faceted exploration, and query suggestion. These applications prove the importance of the proposed TSC problem in many web mining tasks.

The rest of the paper is organized as follows. The related work is given in Section 2. In Section 3, the problem of TSC is formally defined and a series of definitions is given. In Section 4, how to generate the TSC of a concept by web search engines is introduced. Our experiments and results are discussed in Section 5. In Section 6, three applications using the proposed TSC method are introduced. Finally, some conclusions are given.

2. Related work

To the best of our knowledge, the problem of temporal semantic context has not been well studied in existing work. In this section, the related work of the proposed method is given: the recent work of semantic annotations and temporal context.

In the semantic annotation field, with the explosion of community contributed multimedia content available online, many social media repositories (e.g., Flickr,[6] YouTube, and Zooomr[7]) allow users to upload media data and annotate content with descriptive keywords which are called social tags. These tags can be seen as a type of semantic context of the objects such as images or videos. Considering usage patterns and semantic values of social tags, Golder (Golder and Huberman, 2006) mined usage patterns of social tags based on the delicious dataset.[8] Davis (Al-Khalifa and Davis, 2006) concluded that social tags were semantically richer than automatically extracted keywords. Suchanek (Suchanek et al., 2008) used YAGO and WordNet to check

[3] en.wikipedia.org/wiki.
[4] www.linkedin.com.
[5] wordnet.princeton.edu.
[6] www.flickr.com.
[7] www.zooomr.com.
[8] www.delicious.com.