Expert Systems with Applications 41 (2014) 331–341

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Exploiting temporal information in Web search

Sheng Lin [a], Peiquan Jin [a,*], Xujian Zhao [b], Lihua Yue [a]

[a] School of Computer Science and Technology, University of Science and Technology of China, PR China
[b] School of Computer Science and Technology, Southwest University of Science and Technology, PR China

ARTICLE INFO

Keywords:
Temporal expressions extraction
Focused time
Web search
Time-sensitive ranking

ABSTRACT

Time plays important roles in Web search, because most Web pages contain temporal information and a lot of Web queries are time-related. How to integrate temporal information in Web search engines has been a research focus in recent years. However, traditional search engines have little support in processing temporal-textual Web queries. Aiming at solving this problem, in this paper, we concentrate on the extraction of the focused time for Web pages, which refers to the most appropriate time associated with Web pages, and then we used focused time to improve the search efficiency for time-sensitive queries. In particular, three critical issues are deeply studied in this paper. The first issue is to extract implicit temporal expressions from Web pages. The second one is to determine the focused time among all the extracted temporal information, and the last issue is to integrate focused time into a search engine. For the first issue, we propose a new dynamic approach to resolve the implicit temporal expressions in Web pages. For the second issue, we present a score model to determine the focused time for Web pages. Our score model takes into account both the frequency of temporal information in Web pages and the containment relationship among temporal information. For the third issue, we combine the textual similarity and the temporal similarity between queries and documents in the ranking process. To evaluate the effectiveness and efficiency of the proposed approaches, we build a prototype system called Time-Aware Search Engine (TASE). TASE is able to extract both the explicit and implicit temporal expressions for Web pages, and calculate the relevant score between Web pages and each temporal expression, and re-rank search results based on the temporal-textual relevance between Web pages and queries. Finally, we conduct experiments on real data sets. The results show that our approach has high accuracy in resolving implicit temporal expressions and extracting focused time, and has better ranking effectiveness for time-sensitive Web queries than its competitor algorithms.

1. Introduction

Temporal information plays an important role in many research areas such as information extraction, topic detection, question answering, query log analysis, and Web search. Temporal information usually appears in Web pages as temporal expressions, which are typically divided into two types, namely explicit expressions, e.g., March 7, 2012, and implicit expressions, e.g., Today. The various forms of temporal expressions and a mass of temporal information in the Web pages impose some challenging issues within the scope of Web search:

(a) How to determine the right temporal information for implicit expressions contained in Web pages? Differing from the explicit expressions, which can be directly found in a calendar, the implicit expressions need a transformation process and usually a referential time is required.
(b) How to determine the focused time for a Web page? A Web page may contain a lot of temporal information, but which ones are the most appropriate times associated with the Web page? This is very important to temporal-textual Web search engines which support both terms-based and time-based queries, as they aim at finding “the Web pages associated with the given terms and under the given temporal predicate”. For instance, to answering the query specifying “finding the information about tourism during the National Day”, the search engines have to first determine which Web pages are mostly related with “the National Day”.
(c) How to integrate the temporal information in a document into the Web search? As mentioned in Berberich, Bedathur, Alonso, and Weikum (2010), an analysis of Web user query

[*] Corresponding author. Address: School of Computer Science and Technology of China, University of Science and Technology of China, Jinzhai Road 96, Hefei 230027, PR China. Tel.: +86 139 5516 2813.
E-mail address: jpq@ustc.edu.cn (P. Jin).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.048

logs shows that 1.5% of queries are explicitly provided with temporal criteria (Nunes, Ribeiro, & David, 2008), i.e., containing temporal expressions, while about 7% of Web queries have temporal intent implicitly provided (Metzler, Jones, Peng, & Zhang, 2009). So it is worthwhile to find an efficiency way to deal with the time-sensitive queries.

For the first issue, namely implicit time resolution, the difficult part is to select the referential time which is used to resolve implicit expressions. For example, to determine the exact time of the implicit expression "Yesterday" in a Web page, we must know the date of NOW under the context.

For the second issue, namely focused time determination, the difficult part is to develop an effective scoring technique to measure the importance and relevance of the extracted temporal information. As there may be some containment relationship among temporal information, the time ranking task has to consider both frequency and the temporal containment. For instance, suppose "April, 2011" and "17 April, 2011" are two extracted time words, "17 April, 2011" is contained in "April, 2011". Therefore, even "April, 2011" rarely appears in the Web pages, it will still be the focused time for the page in case that there are a great number of extracted time words contained by "April, 2011".

For the third issue, namely time-sensitive ranking, the difficult part is to find a more precise way to calculate the temporal similarity. As the query and the document may contain some temporal expressions, and they may have different representations, the temporal similarity calculation should consider each temporal expression pair between the query and the document, and take the uncertainty of time information into account.

In this paper, we focus on the above three issues and aim to propose effective solutions to the resolution of implicit expressions and the extraction of the focused time for Web pages. The main contributions of the paper can be summarized as follows:

(a) We propose a new dynamic approach to resolve the implicit temporal expressions in Web pages. We classify the implicit expressions into global and local temporal expressions, and then use different method to determine the referential time for global expressions and local expressions (see Section 3).
(b) We present a score model to determine the focused time for Web pages. Our score model takes into account both the frequency of temporal information in Web pages and the containment relationship among temporal information (see Section 4).
(c) We present a new ranking algorithm for temporal-textual Web search. The new algorithm combines the textual similarity and the temporal similarity between queries and Web pages in the ranking process, and can improve search efficiency for time-sensitive queries (see Section 5).
(d) We conduct experiments on real data sets, namely one small data set consisting of 3148 Chinese news articles and a large data set containing 1,812,933 English news articles, to measure the performance of our algorithms. The results showed that our approach got a high accuracy ratio in resolving implicit temporal expressions and extracting focused time, and it also showed that TASE could improve the effectiveness for time-sensitive Web queries and outperforms the competitor algorithms (see Section 6).

2. Related work

In this section, we present an overview of previous research on temporal expressions extraction and time-sensitive ranking methods.

2.1. Temporal expressions extraction

Information about time was first appeared in MUC-5 for information capturing regarding when a joint venture took place. In MUC-6 some research was done on extracting absolute time information as part of general tasks of named entity recognition (Sundheim & Chinchor, 1995). In MUC-7, the notion of temporal information extraction was expanded to include relative time in named entities (Chinchor & Marsh, 1998). MUC is practically the pioneer and prime driver of temporal information extraction research.

The extraction of temporal expressions from documents can be accomplished using an approach similar to named-entity extraction. In general, the first step is to extract time metadata from the document. This can be the creation or last modified date of a file. In case of a Web page, we rely on the information provided by the Web server. The second step is to run a part of speech tagger (POS tagger) on every document. A POS tagger returns the document with parts of speech assigned to each word like noun, verb, etc. The third step is to run a temporal expression tagger like GUTime (Mani & Wilson, 2000) on the POS-tagged version of the document, and this step extracts temporal expressions based on the TimeML standard which has emerged as the standard markup language for events and temporal expressions in natural language (Pustejovsky et al., 2003).

GUTime is part of the TARSQI (Temporal Awareness and Reasoning Systems for Question Interpretation) toolkit (TTK) (Verhagen & Pustejovsky, 2008), which is the state-of-the-art tool for this natural-language processing task. It is based on the TempEx tagger, which is the first temporal tagger for the extraction of temporal expressions and their normalizations (Mani & Wilson, 2000). GUTime has a good performance in the extraction of explicit temporal expressions, but it does not perform very well in dealing with the implicit temporal expressions, especially in the case of lack of the document publication time. To improve the GUTime performance, we need to improve the reference choosing mechanism of GUTime.

Most of the works on temporal expression normalization do not give an effective reference time choosing method for implicit times in real texts. More specifically, the pioneer work by Lascarides, Asher, and Oberlander (1992) investigated various contextual effects on different temporal-reference relations. Then Hitzeman, Moens, and Grover (1995) discussed the reference-choosing taking into account the effects of tense, aspect, temporal adverbials and rhetorical relations. Dorr and Gasterland (2002) presented the enhanced one in addition considering the connecting words. But they are theoretical in nature and heavily dependent on languages. Currently, the static choosing-rules mechanisms (Jang, Baldwin, & Mani, 2004; Lin, Cao, & Yuan, 2008; Vazov, 2001) for reference time choosing are applied into some systems widely. Nevertheless, they are not adaptable to universal implicit times. Zhao, Jin, and Yue (2010) proposed a novel reference time dynamic-choosing mechanism which considers the Global Reference Time and Local Reference Time respectively, and it gets an acceptable performance.

In general, the central role of time in any information space has been studied not only in the area of information extraction, but also in other areas related to information retrieval, such as question answering and summarization (Wong, Xia, Li, & Yuan, 2005). A discussion of different document search and exploration tasks focusing on temporal information embedded in documents is given in (Alonso, Gertz, & Baeza-Yates, 2007).

Sometimes, we need to know the most relevant time for an article, e.g. improving the search result. In general, we can use the frequency of the temporal expressions to determine which the most relevant time is, but it does not take the relation among temporal expressions into consideration. There have been some studies on