Constructing Personal Knowledge Base: Automatic Key-Phrase Extraction from Multiple-Domain Web Pages

Yin-Fu Huang and Cin-Siang Ciou
National Yunlin University of Science and Technology, 123 University Road, Section 3, Touliu, Yunlin, Taiwan 640 huangyf@yuntech.edu.tw

Abstract. In the paper, we proposed a general framework that could automatically extract key-phrases from a collection of web pages concerning a specific topic with the help of The Free Dictionary and then construct a personal knowledge base. Both the base and visual feature in a web page are used to calculate the weight of each candidate phrase. The system extracts top p% key-phrases for each web page based on these two features and then generates a term set using union operators. Next, the system builds the relationships between terms in the term set by referencing The Free Dictionary, and then generates a list of terms sorted by weights. With the top q terms specified by users, a semantic graph can be constructed to present the part of a personal knowledge base, which shows the relationships between terms from the same domain. Finally, the experimental results show that the key-phrases generated by the proposed extractor are with good quality and acceptable for humans.

Keywords: key-phrase extraction, semantic graph, learning mechanism, term correlation, POS.

1 Introduction

With the rapid development and growth of digital information, everyone can create and share electronic documents easily on the Internet. In general, a searching website is organized into different domains such as music, comics, cars, etc. Each domain includes many articles concerning a specific topic. However, it is still not efficient for users to retrieve information related to their preferences, which is an important topic in behavior informatics [12]. If each domain is represented with some key-phrases, it would be much convenient for users to understand the corresponding domain and then determine what they want. Although key-phrases are useful in information retrieval, it is not practical to manually mark them in a collection of web pages because it is a time consuming and tremendous work. Therefore, developing a key-phrase extraction system that can automatically extract key-phrases from a collection of web pages is necessary.
Till now, several key-phrase extraction systems [1-9] have been proposed based on different techniques: supervised algorithms [3, 4, 6, 8, 9] or unsupervised algorithms [2, 5]. A supervised algorithm is to classify a candidate phrase into key-phrase or non-key-phrase after the training, by using a corpus of documents with corresponding author-assigned key-phrases. On the other hand, an unsupervised algorithm usually assigns a numeric score to candidate phrases by considering various features.

Besides, most previous work focused on key-phrase extraction from a document. In this paper, we propose a novel approach which extracts key-phrases from a domain. A domain consists of a collection of web pages concerning a specific topic, from which the proposed approach can generate a list of key-phrases. Moreover, it could work on various domains. First, the system extracts top p% key-phrases for each web page, then collects all the key-phrases from all web pages of the same domain, and finally generates a term set using union operators. Next, the system builds the relationships between terms in the term set by referencing The Free Dictionary, and then generates a list of terms sorted by weights. Finally, with the top q terms specified by users, a semantic graph can be constructed and shows the relationships between terms from the same domain.

The remainder of the paper is organized as follows. A novel approach for key-phrase extraction is proposed in Section 2. Section 3 discusses the experimental results. Finally, we make conclusions in Section 4.

2 System Framework

In this section, we propose the personal knowledge base construction framework as shown in Fig. 1, which consists of the following 8 components: preprocessor, candidate phrase extractor, feature calculation, refinement, correlation matrix generator, term ranking, semantic graph constructor, learning mechanism. Each component is explained in detail in the following subsections.

2.1 Preprocessor

The preprocessor parses web pages and records each term appearing in particular tags, e.g. <title>, <b>, <h1> to <h6>, <i>, <a>, and <meta keywords>. The recorded information would be used as a reference for weighting terms and determining how important candidate phrases are. If a candidate phrase appears in particular tags, it will be given more weight than those not appearing. After parsing, the preprocessor transforms web pages into plain texts; in other words, it removes HTML tags, scripts, etc., and only text blocks are preserved. Then, we employ a state-of-the-art part-of-speech (i.e., POS) tagger [10] to annotate words in the plain texts automatically with part-of-speech tags. Therefore, the preprocessor records each word and its corresponding POS tag. This tagging step is important and necessary because we use linguistic information to help candidate phrase selection.