Mining Unstructured Data Using Field Association Knowledge

Tshering Cigay Dorji, Department of Information Science and Intelligent Systems Graduate School of Advanced Technology and Science, The University of Tokushima, Japan

Abstract

With the tremendous growth in the volume of unstructured textual data on the Internet and company-wide intranets, there is a growing need for automated techniques to “efficiently organize, classify, label, and extract relevant information” (Berry, 2004; Berry & Castellanos, 2008). Merrill Lynch in 1998 cited estimates that as much as 80% of all potentially usable business information originates in unstructured form (Shilakes & Tylman, 1998). Therefore, extracting interesting and non-trivial patterns or knowledge from unstructured data has assumed great importance in fields ranging from business to engineering and biomedical researches (Ramakrishnan, 2009).

Existing techniques may be broadly divided into those based on knowledge engineering approach (Hayes & Weinstein, 1990) and those based on machine learning approach (Pang & Kasabov, 2009; Peng et al., 2008; Graham-Cumming, 2005). Because of the huge amount of time and expertise required to create and maintain knowledge encoding rules for a knowledge engineering approach, the techniques based on machine learning are predominantly used. However, the machine learning techniques also face various challenges such as the problem of high dimensionality, and selecting features that really represent documents well without the loss of semantic information.

Therefore, this thesis presents a more effective technique for mining unstructured textual data using field association knowledge. The field association knowledge base is represented by a comprehensive dictionary of Field Association (FA) Terms (Tsuji et al., 1999; Fuketa et al., 2000; Atlam et al., 2006). FA Terms are those single (uni-word) and compound (multi-word) terms that have high field specificity and can discriminate between fields. These FA Terms are connected to a hierarchical knowledge organization scheme called a field tree (Fuketa et al., 2000).

The construction of the field association knowledge base is basically a one-time process and also does not involve complex knowledge encoding rules, though it may need some regular updating to keep its contents up-to-date. Once the knowledge base is constructed, it can be used for mining tasks in all domains covered by it. But the first problem lies in the lack of an effective method to extract and select relevant FA Terms to build such a knowledge base. So, firstly, this thesis presents a new method to extract, select and rank FA Terms from domain-specific corpora using part-of-speech (POS) pattern rules, corpora comparison and modified \textit{tf-idf} weighting. Experimental evaluation on 21 fields using 306 MB of domain-specific corpora obtained from English Wikipedia dumps selected a total of 31,234 FA Terms, with up to 2,517 FA Terms (single and compound) per field at precision and recall of 75-97% and 65-98% respectively. Compared to the traditional methods, this method selects higher number of relevant FA Terms at high precision and recall.
This thesis then presents the methodologies for using the field association knowledge base to achieve four major tasks of mining unstructured data, viz. text classification, document summarization, document clustering and information extraction. Since mining unstructured data using field association knowledge base is essentially a classification problem, the new methodology is compared against Naïve Bayes, $k$-Nearest Neighbour (kNN) and Centroid-based classifiers to classify 9,449 documents from Reuters RCV1 Corpus, 20-Newsgroup and 4-Universities data sets. The new methodology outperformed others with a precision of 97.50% followed by Centroid-based with 84.67%, Naïve Bayes with 78.11% and kNN with 47.66%.

Thirdly, in order to store the field association knowledge base efficiently in memory, three new methods to compress a dynamic trie data structure using Minimal Prefix (MP) double array are presented. Compared to a Ternary Search Tree which is a popular implementation of trie often used as the benchmark against which the performance of new implementations of dynamic trie are measured (Clement et al., 1998; Heinz et al., 2002), the key retrieval of the compressed MP double array trie is 50% faster and its size is 3-5 times smaller.