**Multi-dimensional Text Warehousing for Text Analytics**

**Jiyun Kim**, **Han-joon Kim**

1School of Electrical and Computer Engineering, University of Seoul, Seoul, Korea

**Abstract** - A data warehouse is a repository of integrated data that provides the basis for decision making which is required to establish crucial business strategies. Recent data warehouses focus upon not only structured data but also textual data. Specially, a text warehouse is a type of data warehouse that provides efficient document retrieval and summarization capabilities. In this paper, we propose a novel way of text warehousing to support text mining capabilities beyond the simple search and aggregation functions. The proposed text warehousing method allows us to perform text mining tasks such as text classification, text clustering, and word association mining more effectively.

**Keywords:** Data warehouse, Text warehouse, Text mining, OLAP

1 **Introduction**

Nowadays, a data warehouse that provides multi-dimensional OLAP capabilities is an important information repository for business intelligence. It can provide significant multi-dimensional views of target measures by integrating huge amounts of structured data retrieved from various sources of relational databases. Currently, due to explosive growth of textual data, many studies on text (or document) warehouses have been carried out to develop efficient document retrieval and summarization capabilities.

In this paper, we introduce a new way of building a text warehouse that provides text mining capabilities as well as simple search and aggregation functions. Our proposed method makes it possible to perform various text mining tasks such as text classification, text clustering, and word association mining. In order to support more reliable text mining tasks, we intend to extract concepts hidden in documents or words. For this, we employ the Wikipedia articles. The topological relationships among concepts are determined in our previous work using the search engine named **Elasticsearch**. Significant probabilistic weights for text mining are stored in the text warehouse, and they are used in conducting text mining algorithms such as naïve Bayes classification, EM clustering, and A priori-based word associations.

2 **Backgrounds**

Wikipedia is a free-content Internet encyclopedia, supported by the non-profit Wikimedia Foundation [1]. Each of Wikipedia articles allows defining a concept [2]. A Wikipedia article includes a title and a body text, and the body text contains an info-box and an anchor texts. These components are used in isolating good quality concept-level articles. In [3], G. Lee proposed a method of building corpus-dependent topic graphs using Wikipedia-based **Elasticsearch** search engine. With a relational database containing well-chosen Wikipedia articles, we can develop the concept network that depends on a given document corpus, and the concept network is materialized as a dimension table in our text warehouse.

3 **Text Warehousing for Text Mining**

3.1 **Text warehouses**

Our goal is to develop a text warehouse (TW) that provides text mining capabilities as well as simple search and aggregation functions. Figure 1 shows the proposed architecture for the proposed text warehouse.

![Figure 1. System architecture of the proposed text warehouse](image_url)

Similarly to a general data warehousing, we firstly perform integrating document data from diverse data sources, and conduct pre-processing tasks such as keyword extraction, tokenizing, semantic tagging and document summarization. At this time, we record the metadata such as title, author, subject, date, and format. Our text warehouse is developed according to the multi-dimensional schema given in Figure 2. The important thing is that the sense of words occurring in incoming documents is saved with its corresponding weights in TW, and also the information of relevant words are saved in a dimension table of TW by using WordNet; WordNet is a lexical database for the English language, which consists of synonyms, short definitions and usage examples [4]. The developed text warehouse will be used for OLAP operations (such as drill-down and roll-up), and text mining tasks (such as text classification and clustering).
3.2 Multi-dimensional modeling

Figure 2 shows the database schema of the TW that supports text mining tasks; the schema consists of two fact tables, several dimension tables and bridge tables. The ‘document’ fact table stores the documents collected and their metadata. The ‘word’ fact table stores the word information of the document collected. The ‘document word’ bridge table is located between ‘document’ fact and ‘word’ fact. These tables have a many-to-many relationship with each other, and the bridge table yields a one-to-many relationship with two facts [5].

The ‘category’ dimension table expresses the category of documents (or words). The ‘part of speech’ dimension table contains the information what type of word it is. The concepts extracted from the Wikipedia database are stored in the ‘concept’ dimension table, which has topological relationships among concepts. The keywords extracted from the WordNet database are stored in the ‘keyword’ dimension table, which has the metadata about keywords and their synonyms. Because each of dimension tables and the fact table are the many-to-many relationships, the bridge table is required between the tables.

3.3 Weight factors for text mining

As stated before, our proposed TW supports text mining tasks, and for this various types of weights are stored in the bridge table. In our work, the weights are computed in a probabilistic manner. As for text classification, Bayes’ theorem can be used to find the maximum posterior probability, and then the probability that the document d is classified into the category c is simplified as follows:

\[ P(c|d) \approx \prod_{w} P(w_i|c)P(c) \]  

(1)

where \( w_i \) denotes each of words occurring in the document d.

For more accurate text mining tasks, we consider the sense (or meaning) of words. Thus Equation 1 is expanded as Equation 2 with the additional random variable; i.e., the concept (or sense) \( s_j \) that the word \( w_j \) has.

\[ P(c|d) \approx \prod_{w} \sum_{s_j} P(w_i|s_j, c) P(s_j|c)P(c) \]  

(2)

where \( s_j \) denotes a concept extracted from Wikipedia. \( P(w_i|s_j, c) \) corresponds to the ‘WordConceptWeight’ field of ‘Word_Concept’ bridge table, and \( P(c) \) corresponds to the ‘DocCategoryWeight’ field of the ‘Doc_Category’ bridge table. These weights significantly contribute to enhance text classification algorithms (such as naïve Bayes) and text clustering algorithms (such as EM).

4 Summary

In this paper, we have suggested a new way of building the text warehouses to support text mining tasks. Our proposed method emphasizes the sense of words occurring in documents, and thus all of related probabilistic weights are stored in the bridge tables. Here, the word sense is accounted for in the concept space defined with the Wikipedia. Therefore, we expect that our text warehouse makes it possible to perform most of text mining tasks more efficiently.

5 Acknowledgments

This research was supported by a grant (15AUDP-B100356-01) from Urban Architecture Research Program funded by Ministry of Land, Infrastructure and Transport of Korean government, and was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF-2015R1D1A1A09061299) funded by the Ministry of Education.

6 References