Chapter 8
System Analysis and Design for Document Classification

ABSTRACT

The text-mining process starts with a keyword search in text collections. Current text processing technology allows a search technique beyond simple Boolean searches by using natural language queries. Since search engines can recognize any of thousands of keywords and phrases but not the concepts behind the text, it is necessary for researchers to construct an automatic keyword extractor to generate the "Keyword List" for each document. Later, this list can act as the knowledge base to associate unorganized documents to meaningful classes. Failures in identifying the keywords for a certain concept will result in missing values or data for that specific concept.

FACT FINDING

Technology is nearly at a point that Text Classification can apply automated **Text Mining** approaches to develop strategic information. Much of the relevant information is contained on Textual documents and is freely available if one can gain access to it. **Text Mining** applications are expensive and relatively crude, but as interest grows in this, prices will diminish and functionality will improve. Domain intelligence plays an important role when **Text Mining** tools make strategic and operational decisions.

One particular focus for IT has been on using Data Mining techniques to extract meaningful patterns and build predictive customer relationship

DOI: 10.4018/978-1-7998-3772-5.ch008

models from textual data. Although widely used, data mining is currently widely available only to structured, numeric databases. However, a majority of business information exists in the form of unstructured or semi structured text documents or in Web based data sources. The traditional way of processing text information involves human actions in information gathering, analysis, and dissemination. This requires substantial investment of money, time, and human resources.

Moreover, it is difficult to combine qualitative text data with quantitative numeric data in business analyses. Therefore, there is a pressing need to develop a method that can accurately extract business intelligence from large text collections and integrate the fragmented information into business intelligence databases.

The text-mining process starts with a keyword search in text collections. Current text processing technology allows a search technique beyond simple Boolean searches by using natural language queries. Since search engines can recognize any of thousands of keywords and phrases but not the concepts behind the text, it is necessary for researchers to construct an automatic keyword extractor to generate the "Keyword List" for each Document. Later, this list can act as the knowledge base to associate unorganized Documents to meaningful Classes. Failures in identifying the keywords for a certain concept will result in missing values or data for that specific concept.