Abstract

Text mining is the analysis of unstructured data by combining techniques from knowledge discovery in databases, natural language processing, information retrieval, and machine learning. Text mining allows us to analyze web content dynamically to find meaningful patterns within large collections of textual data. There are too many economic news articles to read. Therefore, it is necessary to summarize them. In this study, TM is used to analyze the vast amount of text produced in newspaper articles in Turkey. We mine unstructured economy news with natural language processing techniques including tokenization, transform cases, filtering stopwords and stemming. Similarity analysis is also used to determine similar documents. The word vector is extracted. Therefore, economy news is structured into numeric representations that summarize them. In addition, k-means clustering is used. Consequently, the clusters and similarities of the articles are obtained.

Keywords: Knowledge discovery in databases, Text mining, Natural language processing

1. Introduction

With the rapid growth of the Internet, the digital platform is replacing traditional media. A huge amount of big data are available in the digital platform. In addition, real-world data come in a variety of forms and can also be extremely large. Data mining (DM) technology is used to extract patterns from well-structured data that exist in databases. However, 80% of these data exist as textual form. Therefore, it becomes necessary to use text mining (TM) techniques. TM, which is a very complex process, draws on data analysis techniques to find meaningful patterns within these large collections of unstructured data.

* Corresponding author.
E-mail address: esrakahya@erciyes.edu.tr (E. Kahya)
Identification of useful information from textual databases through the application of different DM techniques began in the early 2000s and has been widely used in websites (Glancy, Yadav, 2011; Thorleuchter, Van den Poel, 2012; Soibelman, Wu, et al. 2008; Wong, Lam, 2009, Rao, et al. 2014). Engel, et al. (2010) developed algorithms to find and characterize changes in topic within text streams. They monitored a stream of text for changes within the content of that data stream. Hu and Liu (2012) presented one real-world application to further illustrate how to utilize text analytics to solve problems in social media applications. They introduced an effective way to improve recently short text representation quality by integrating semantic knowledge resources. Choi and Kim (2013) proposed a social relation extraction system based on the dependency kernels of SVMs. Spaeth and Desmarais (2013) recently explored TM to improve classical collaborative filtering methods for a site aimed at matching people who are looking for expert advice on a specific topic. They compared results from an LSA-based text similarity analysis, a simple user-user collaborative filter, and a combination of both methods used to suggest people meet for a knowledge-sharing WS. Kahya-Ozyirmidokuz (2013) used TM to analyze online Turkish social shopping firms. Text preprocessing techniques (tokenization, term filtering methods, Euclidean distance measure etc.) were used. Kahya-Ozyirmidokuz and Ozyirmidokuz (2014) analyzed the top seven heating systems firms in Turkey in terms of customers’ complaints. They not only grouped the customer complaints in the heating sector but also the firms to which complaints were made.

Electronics newspapers have gradually become the main sources for news readers. When facing numerous reports on a series of events in a topic, a summary of stories from news reports will benefit news readers in reviewing the news topic efficiently. Besides identifying events and presenting news titles and keywords that the topic detection and tracking techniques are used to do, a summarized text to present event evolution is necessary for general news readers to review events under a news topic. (Lin, Liang, 2008). There are too many economic news and financial reports to read. TM uses information retrieval, natural language processing, information extraction and DM to summarize them. In recent years there have been a lot of studies which used newspapers (Hagenau, et al. 2013; Li, Ng, 2013; Geva, Zahavi, 2014). Many studies have been conducted which analyzed unstructured news data (Bai, 2010; Cecchini et al., 2010, Groth, Muntermann, 2011; Chan, Franklin, 2011; Zhang, et al. 2012; Fortuny et al., 2014). Ong et al. (2005) focused on the automatic generation of a hierarchical knowledge map, NewsMap, which is based on online Chinese news, particularly the finance and health sections.

Lin and Liang (2008) proposed a topic retrospection process and implemented the SToRe (Story-line based Topic Retrospection) system that identifies various events under a news topic, and composes a summary by means which news readers can get an outline of event evolution on a particular topic. Huang et al. (2010) proposed a financial news headline agent to assist investors in deciding to buy or sell stocks in the Taiwan stock market after receiving the essential real-time news headline disseminated by the agent.

Chen (2010) presented an intelligent location-based service system to provide local news with summaries to a personal handheld PDA (Personal Digital Assistant) device based on the user’s location. Ma, Pant and Sheng (2011) presented an approach that uses graph-theoretic measures and machine learning techniques to infer competitor relationships on the basis of the structure of an intercompany network derived from company citations (co-occurrence) in online news articles. Fortuny et al. (2012) gathered and analyzed 68,000 related online news articles published in 2011 in Flemish newspapers. These articles were analyzed by a custom-built expert system. Yu et al. (2013) proposed the presence and intensity of emotion words as features to classify the sentiment of stock market news articles. Chung (2014) developed BizPro, an intelligent system for extracting and categorizing BI factors from news articles.

The aim of this research is to extract knowledge within 10 Turkish newspapers’ economy news articles which are unstructured and fuzzy. We analyze the similarities of ten Turkish newspapers’ economy news ranging from 01 March 2014 to 10 March 2014. Five hundred and forty three documents are used. The paper is organized as follows. The TM process is introduced in Section 2. Section 3 gives the details of the application. The conclusions are drawn in Section 4.

2. Text Mining Process

The text is a written language in the digital platform. The amount of existing text today is ever growing. Classical DM techniques are not sufficient to analyze unstructured data. We need to use some linguistic approaches. TM
processes unstructured information, extracts meaningful numeric indices from the text, and, thus, makes the information contained in the text accessible to the various DM algorithms, including statistical and machine learning. Businesses use DM and TM to analyze customer and competitor data to improve competitiveness. The benefits of TM are obvious in areas where a large number of textual data are collected from business transactions. For example, the free-form text of customer interactions allows trending during time in the areas of complaint (and praise), warranty claims and error tracking, all of which are clearly input for product development and service allocation (Feldman and Sanger, 2007). Figure 1 (Chakraborty, et al., 2013) illustrates the TM process.

After data collection, the next step is to extract, clean, and create a dictionary of words from the documents using natural language processing. This includes identifying sentences, determining parts of speech, and stemming words. This step involves parsing the extracted words to identify entities, removing stopwords, and spell-checking. In addition to extracting words from documents, variables associated with the text such as date, author, gender, category, etc., are retrieved. The most important task after parsing is text transformation. This step deals with the numerical representation of the text using linear algebra-based methods, such as latent semantic analysis (LSA), latent semantic indexing (LSI), and vector space model. This exercise results in the creation of a term-by-document matrix (a spreadsheet or flat-like numeric representation of textual data). The dimensions of the matrix are determined by the number of documents and the number of terms in the collection. This step might involve dimension reduction of the term-by-document matrix using singular value decomposition (SVD). In a corpus of several thousands of documents, you will likely have many terms that are irrelevant to either differentiating documents from each other or to summarizing the documents. You will have to manually browse through the terms to eliminate irrelevant terms. This is often one of the most time-consuming and subjective tasks in all of the TM steps. It requires a fair amount of subject matter knowledge (or domain expertise). In addition to term filtering, documents irrelevant to the analysis are searched using keywords. Documents are filtered if they do not contain some of the terms or filtered based on one of the other document variables such as date, category, etc. Term filtering or document filtering alters the term-by-document matrix. The term-by-document matrix contains the frequency of the occurrence of the term in the document as the presence of the term in a document as the value for each cell. From this frequency matrix, a matrix is generated using various term-weighting techniques. The TM step involves applying traditional DM algorithms such as clustering, classification, association analysis, and link analysis. TM is an iterative process, which involves repeating the analysis using different settings and including or excluding terms for better results. The outcome of this step can be clusters of documents, lists of single-term or multi-term topics, or rules that answer a classification problem (Chakraborty, Pagolu and Garla, 2013).

Figure 1: Text mining process
Source: (Chakraborty, Pagolu and Garla, 2013)
The document processing techniques which are used in the study are as follows (Ingersoll, 2013):

- **Tokenization**: The process of breaking up a string into tokens to be indexed. Proper, consistent handling of punctuation, numbers, and other symbols is important. For instance, tokenizing microprocessor might mean outputting several tokens (micro, processor, and microprocessor) so that user queries for variations are more likely to succeed.
- **Downcasing**: All words are converted to lowercase, making it easy to do a case-insensitive search.
- **Stemming**: Strips words of suffixes, and so on.
- **Stopword removal**: Remove commonly occurring words like the, and, and a that occur in most documents. Originally done to save space in the index, but some newer search engines no longer remove stopwords since they can help in more advanced queries.
- **Synonym expansion**: For each token, synonyms are looked up in a thesaurus and added to the index. This is often done on the query terms instead of the index terms, since updates to the synonym list can be accounted for dynamically at query time without the need to re-index.

Once the text is transformed into a set of numbers that adequately capture the patterns in the textual data, any traditional statistical or forecasting model or DM algorithm can be used on the numbers for generating insights or for predictive modeling (Chakraborty, Pagolu and Garla, 2013). Although there are numerous techniques in TM, the aim is to discover hidden and useful knowledge.

Most existing approaches focus on the manual assignment of keywords by professional curators who may use a fixed taxonomy, or rely on the authors’ judgment to provide a representative list. Research has therefore focused on methods to automatically extract keywords from documents as an aid either to suggest keywords for a professional indexer or to generate summary features for documents that would otherwise be inaccessible (Rose, et al., 2010).

### 2.1. Term Frequency Inverse Document Frequency (TF-IDF)

As seen in Equation (1) and Equation (2), TF-IDF weight determines the importance of a word in a document collection. Term frequency is the number of times a term \( t \) appears in a document \( d \). The inverse document frequency is the measurement for how rare a term is in documents.

\[
\text{Equation (1)} \quad t_f(t,d) \cdot \frac{f(t,d)}{\max \{ f(t,d) : w \cdot d \}} \\
\text{Equation (2)} \quad \text{idf}(t,D) \cdot \log\left( \frac{|D|}{|d \cdot D : t \cdot d|} \right)
\]

### 2.2. Similarity analysis

Clustering is a useful technique that organizes a large quantity of unordered text documents into a small number of meaningful and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanisms. Partitional clustering algorithms have been recognized to be more suitable as opposed to the hierarchical clustering schemes for processing large datasets. A wide variety of distance functions and similarity measures have been used for clustering, such as squared Euclidean distance, cosine similarity, and relative entropy (Huang, 2008).

Before clustering, a similarity/distance measure must be determined. The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. In many cases, these characteristics are dependent on the data or the problem context at hand, and there is no measure that is universally best for all kinds of clustering problems (Huang, 2008).

Although similarity between documents is an essential ingredient in organizing unlabeled documents into distinct groups, measuring the similarity of documents is an end in itself. Measuring similarity between documents is
fundamental to most forms of document analysis, especially information retrieval (Weiss et al., 2005). Given two documents $t_a$ and $t_b$, their cosine similarity is as follows:

$$\text{SIM}_{\cos}(t_a, t_b) = \frac{t_a \cdot t_b}{\|t_a\| \|t_b\|}$$

where $t_a$ and $t_b$ are $m$-dimensional vectors over the term set $T \cup \{t_1, ..., t_m\}$. Each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between $[0,1]$ (Huang, 2008).

3. Application

After the news articles have been gathered automatically they have to be prepared before they can be categorized. Newspaper URLs are transformed into a collection of documents by generating a document for each record. TF-IDF is used for vector creation in the document processing step. Tokenization is applied to break the text up into words. The transform cases step turns all letters into lower case. Then we eliminate stopwords which are useless and redundant including English and Turkish stopwords and HTML codes. The stemming process uses a snowball based stemmer for the Turkish language. A vector representation of the articles is achieved. Table 1 illustrates the top 4 words’ list from the preprocessing.

<table>
<thead>
<tr>
<th>Number</th>
<th>Words in Turkish</th>
<th>Meanings</th>
<th>Total occurrences</th>
<th>Document occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>miliyet</td>
<td>nationality</td>
<td>8026</td>
<td>68</td>
</tr>
<tr>
<td>2</td>
<td>ekonomi</td>
<td>economy</td>
<td>6296</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>hurriyet</td>
<td>liberty</td>
<td>6017</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>radikal</td>
<td>radical</td>
<td>3686</td>
<td>44</td>
</tr>
</tbody>
</table>

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Cosine similarity is one of the most popular similarity measures applied to text documents, such as in numerous information retrieval applications and clustering too (Huang, 2008). The word vector is used in similarity analysis to determine similar documents. Cosine similarity is used in similarity analysis. The histogram of similarity analysis is given in Figure 2. Figure 3 presents the similarity graphs. Some of the similarity distances of the documents are given in Table 2. Figure 4 shows the k-distance plot graph for $k=3$. 
Figure 2: Histogram of similarity analysis

Figure 3: Similarity graphs

Table 2: Examples of similarity analysis

<table>
<thead>
<tr>
<th>News-1</th>
<th>News-2</th>
<th>Similarity distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>396</td>
<td>0.708</td>
</tr>
<tr>
<td>9</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>67</td>
<td>1.401</td>
</tr>
<tr>
<td>29</td>
<td>48</td>
<td>1.399</td>
</tr>
</tbody>
</table>
When clustering is employed on web sites, we are usually more interested in clustering the component pages according to the type of information that is presented in the page (Huang, 2008). Two groups are detected with k-means clustering. The centroid plot view of clusters can be seen from Figure 5.

A performance operator, which can be used to derive a performance measure (in the form of a performance vector) from the dataset, is used. The performance vector of the model’s cluster number index is 0.894.
4. Conclusions

In this study a TM model is developed to extract useful clusters from the economy news of newspapers. Unstructured documents are preprocessed and graphs, tables, keywords and the word vector are achieved. The word vector is used in similarity analysis to determine similar documents. Documents are transposed into numerical values. Therefore, traditional DM techniques can be applied. Alternative TM techniques can be studied in future research to compare various approaches and to implement this framework. The quality of electronic news can also be assessed through TM. The model can be improved in estimation if documents are collected on a particular subject.

References


