International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 3, March-2014, pp: (446-451), Impact Factor: 1.252, Available online at: www.erpublications.com

Text Mining Techniques: A Semantic Approach in order to classify the documents

Dr. Manoj Kumar Singh[1], Mohammad Kemal Ahmad[2], Dr. Mohammad Iqbal[3]

[1]Asst. Professor, Department of Computing, Adama Science & Technology University, Adama, Ethiopia
[2]Head, Department of Computing, Adama Science & Technology University, Adama, Ethiopia
[3]Asst. Professor, Department of Computing, Adama Science & Technology University, Adama, Ethiopia

Abstract: Over the past two decades, the automatic management of electronic documents has been a major research field in computer science. Text documents have become the most common type of information repositories especially with the increased popularity of the internet and the World Wide Web. Internet and web documents like web pages, emails, newsgroup messages, internet news feed etc., contain million or even billion of text documents. In the last decades content-based document management tasks have gained a prominent status in the information systems field, due to the increased availability of documents in digital form. Though the paper suggests techniques for classifying Research papers pertaining to Computer Science.

Keywords: Text Mining, documents, classification, research paper, DSL.

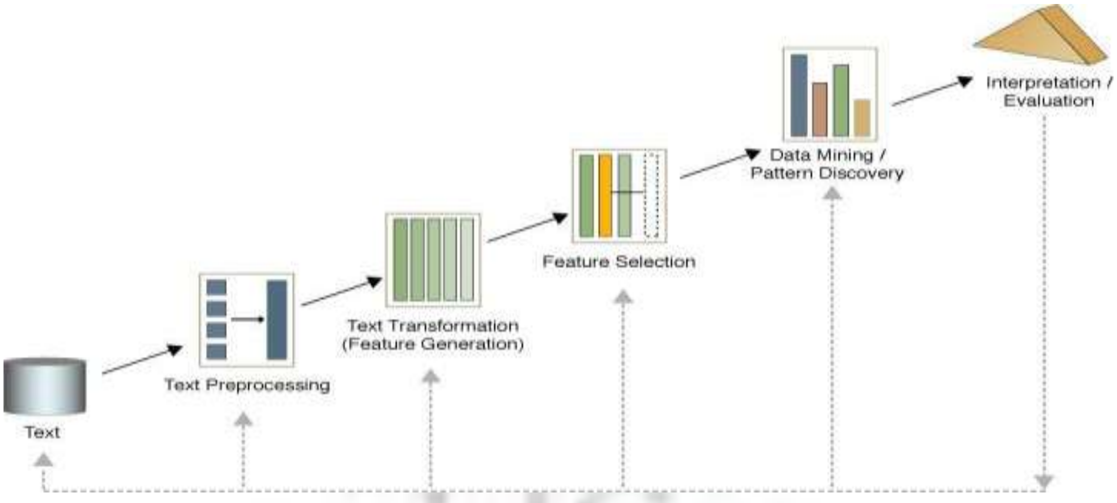
Introduction

Modern information systems allow firms to capture vast amounts of data. Much of this data is structured data that can be analyzed using traditional database software. Increasingly, however, large amounts of data such as textual data are unstructured, and defy simple attempts to make sense of it. Manual analysis of this unstructured textual data is increasingly impractical, and as a result, text mining Techniques are being developed to automate the process of analyzing this data. Text mining has been used to identify intellectual core in the information systems discipline topic discovery customer relationship management and target advertising. Elsewhere text mining has been used not only to identify sentiments of investors such as negative or positive opinions from business news papers and financial websites but also to detect objects from images. Given the need to analyze and understand textual data, and given the increasing popularity of text mining methods, the time seems right for a research essay to examine the major methods available for text mining, and to address the topic of how to appropriately apply text mining techniques in research and practice.

Now days the automatic management of electronic documents has been a major research field in computer science. Text documents have become the most common type of information repositories especially with the increased popularity of the internet and the World Wide Web. Internet and web documents like web pages, emails, newsgroup messages, internet news feed etc., contain million or even billion of text documents. In the last decades content-based document management tasks have gained a prominent status in the information systems field, due to the increased availability of documents in digital form.

Text mining deals the categories of operations, retrieval, classification (supervised, unsupervised and semi supervised) summarization, trend and association analysis. The main goal of text mining is to enable users to extract information from textual resources. How the documented can be proper annotated, presented and classified, so the documents categorization consist several challenges, proper annotation to the documents, appropriate document representation, an appropriate classifier function to obtain good generalization and avoid over-fitting, also an appropriate dimensionality reduction method need to handle algorithmic issues.

Text Mining is the automated or partially automated processing of text. Classifying text documents, analyzing syntax, identifying relationships among documents understanding a question expressed in natural language, extracting meaning from message, summarizing involve application of non trivial tasks on textual data.

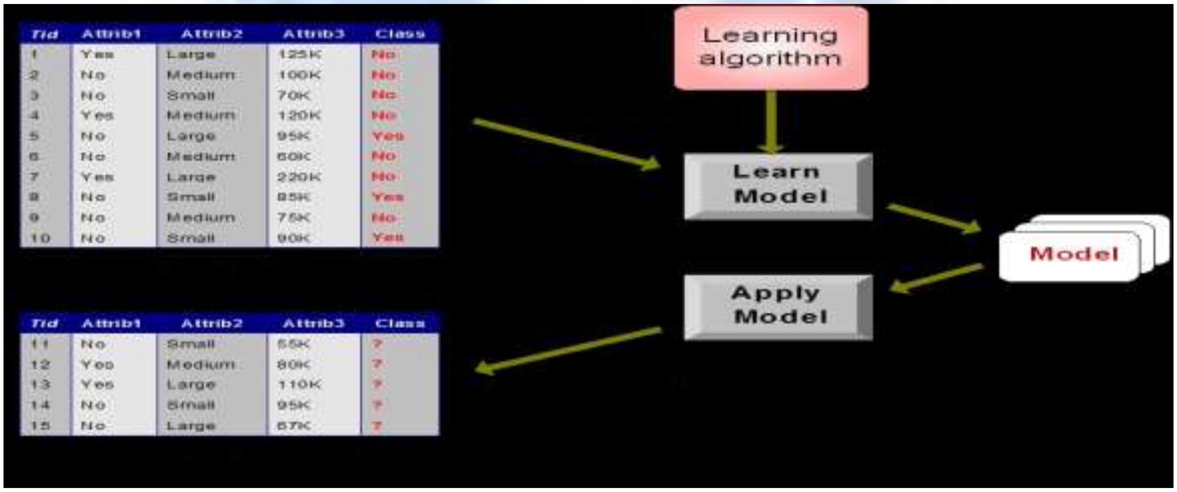


-

Text Mining Process

Classification:

Classification is a data mining (machine learning) technique used to predict group membership for data instances. Classification is used to extract models describing important data classes or to predict future data trends. Classification predicts categorical labels. Data classification is a two step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. Each tuple is assumed to belong to a predefined class, called the class label attribute. In the context of classification, data tuples are also referred to as samples, examples, or objects. The data tuples analyzed to build the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. Since the class label of each training sample is provided, this step is also known as supervised learning. The learned model is represented in the form of classification rules, decision trees, or mathematical formulae. In the second step, the model generated is used for classification.



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

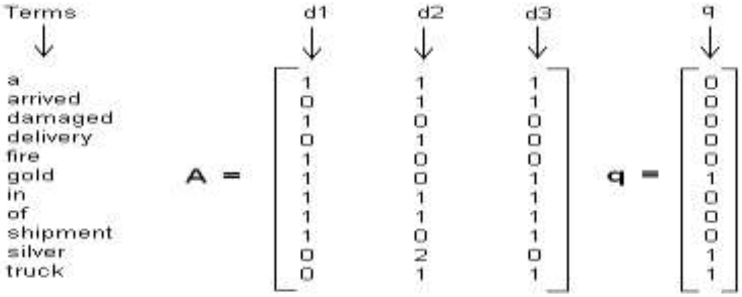
Document Preprocessing

Text documents require preprocessing before mining techniques can be applied.
Preprocessing Techniques include:

1. Stop Word Removal: Many words are not informative and thus irrelevant for document representation For e.g. the, and, a, an, is, of that, etc.
2. Reducing the words to their forms.

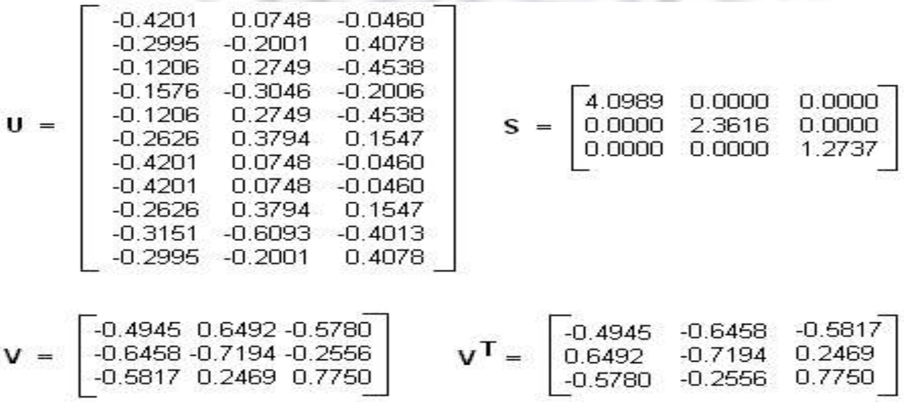
A document may contain occurrences of words like fishes, fisher and fishers (would not be retrieved by a query with the keyword fishing). Different words share the same word stem and should be represented with its stem, instead of the actual word i.e. fish.

Table 1: Illustration of Text Mining Data



$$\begin{array}{l} \text{Terms} \\ \text{a} \\ \text{arrived} \\ \text{damaged} \\ \text{delivery} \\ \text{fire} \\ \text{gold} \\ \text{in} \\ \text{of} \\ \text{shipment} \\ \text{silver} \\ \text{truck} \end{array} \qquad A = \begin{array}{ccc} d1 & d2 & d3 \\ \hline 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{array} \qquad q = \begin{array}{c} q \\ \hline 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{array}$$

Table 2: After Dimensionality reduction (k=2)



$$U = \begin{bmatrix} -0.4201 & 0.0748 & -0.0460 \\ -0.2995 & -0.2001 & 0.4078 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.1576 & -0.3046 & -0.2006 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.3151 & -0.6093 & -0.4013 \\ -0.2995 & -0.2001 & 0.4078 \end{bmatrix} \qquad S = \begin{bmatrix} 4.0989 & 0.0000 & 0.0000 \\ 0.0000 & 2.3616 & 0.0000 \\ 0.0000 & 0.0000 & 1.2737 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.4945 & 0.6492 & -0.5780 \\ -0.6458 & -0.7194 & -0.2556 \\ -0.5817 & 0.2469 & 0.7750 \end{bmatrix} \qquad V^{T} = \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \\ -0.5780 & -0.2556 & 0.7750 \end{bmatrix}$$

The entire task of Classification is proposed as a Framework

The proposed framework provides flexibility to the user and enables him to configure parameters for the experiment. The user can specify and set up parameters relating to:

1. Construction of Domain Specific Lexicon (DSL).
2. Selection of DSL.
3. Building a Corpus of DSL.
4. Building of Corpus of Research Documents used for testing the model.
5. Selecting the format of the file/database for storing the results.
6. Selecting the pathname/filename of the output files/database tables.
7. Building the Concept Matrix for the CPTR model.
8. Building Corpus of Documents for Testing the CPTR model.

Classification Level-1

This is the first step in the Classification framework. The objective of this classification is to derive a classification key which is comprised of

a) The Year (in which the Paper is published)
b) The Journal (in which the Paper is published)

This classification does not involve the usage of Text Mining techniques and is based on tagged parameters captured from the documents or accepted from the user.

Algorithm

Capture /Accept tagged data from the Research document and store the same in a database table along with the generated Classification key. Terms:

Cr - Corpus of Research Documents

$_{Di}$ - Research Document wherein $d_{i \in Cr}$

Cr = {d1,d2,d3…dn}

$_{Td}$= Tagged data from D_i

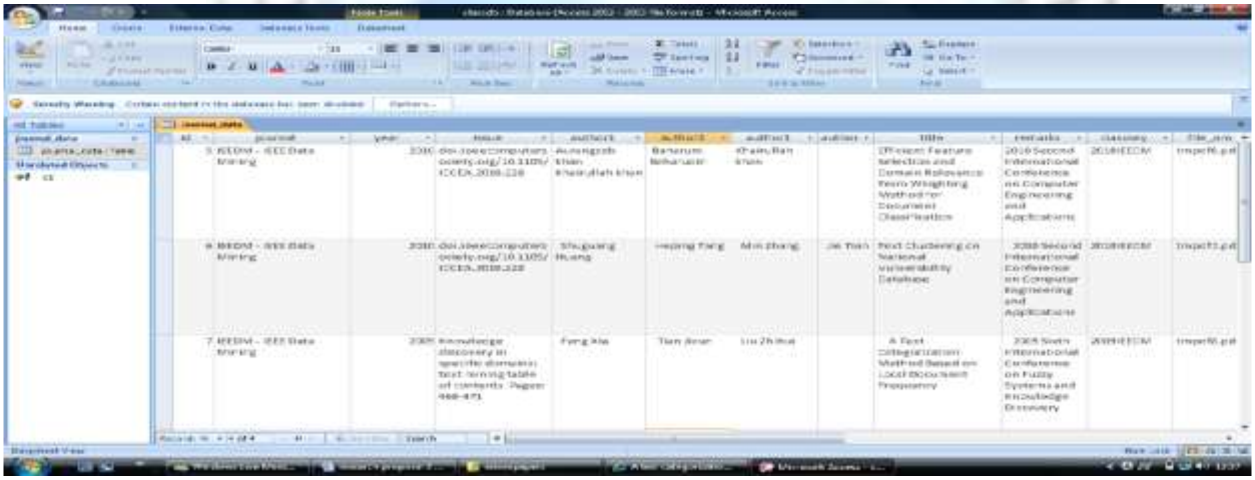
Input: Tk_i= Tagged data from each d_i

Output: Database Record containing Tagged Data + Generated Classification Key for each $D_{i \in Cr.}$

Steps:

```
For each $D_i \in_{Cr}$
Accept tagged data ($D_i$,$Tk_i$) ;
Generate Classification Key = Year of publish + Journal id;
Generate unique primary key;
Insert new
Database records,
End;
```

Capturing tagged Data



ID	journal	year	issue	author1	author2	author3	author4	title	remarks	classkey	file_name
5	IEEDM - IEEE Data Mining	2010	doi.ieeecomputersociety.org/10.1109/ICCEA.2010.228	Aurangzeb khan Khairullah khan	Baharum Baharudin	Khairullah khan		Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification	2010 Second International Conference on Computer Engineering and Applications	2010IEEDM	tmpe[illegible].pdf
6	IEEDM - IEEE Data Mining	2010	doi.ieeecomputersociety.org/10.1109/ICCEA.2010.228	Shuguang Huang	Heping Fang	Min Zhang	Jie Tian	Text Clustering on National Vulnerability Database	2010 Second International Conference on Computer Engineering and Applications	2010IEEDM	tmp[illegible].pdf
7	IEEDM - IEEE Data Mining	2009	Knowledge discovery in specific domains: text mining table of contents Pages 468-471	Feng Xia	Tian [illegible]	Liu Zhihui		A Text Categorization Method Based on Local Document Frequency	2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery	2009IEEDM	tmp[illegible].pdf

S No.	Attributes	Type	comments
1.	Id	Numeric	Primary
2.	Journal	Text	
3.	Tittle	Text	
4.	Year	Numeric	
5.	Issue	Text	
6.	Author1	Text	
7	Author2	Text	

Database table (journal data)- inserted records

Classification - Level 2

The proposed Hierarchical Classification -Level 2 of Research Documents uses text Mining techniques to assign labels to research papers indicating their domain/sub domain or area of research. The classification labels are predicted on the basis of the title and keywords found in the research papers. The Titles and keywords from the research documents are extracted and stored in the form of PDF files. Labels are predicted for this Corpus of PDF files. Assigning/predicting class labels is equivalent to mapping of a Research Document with it predicted class label.
Level -2 Classifications involves 3 major processes:

a) The Training Process
b) The Testing/Application Process
c) Process for Storing Predictions (Results) and Generating key.

(a) The Training Process:

Steps:

1. Identify Domain Specific terms and construct DSL file(s).
2. Associate a Classification label with each of the DSL file(s).
3. Build a Corpus of DSL file(s). Cr = {dsl_1, dsl2, dsl3…dsl_n}
4. Preprocess the Corpus Cr

a) Apply a Stemming algorithm to reduce all the words to their root form.
b) Generate VSM or a Term Document matrix using Binary Term Occurrence

D (i, j) (where i is the document i and j is the jth term of document i.)

(TF and TF-IDF are not used in the matrix because only the occurrence of the term in the DSL file is relevant for classification; the distinguishing or rarity of the term is irrelevant in this approach).
5. Train the K-NN Classifier using C as training examples.

(b) Testing /Application of the Classifier:

The Title of the Research paper and the keywords contained in it are extracted and stored in the separate file each. This file is in the PDF format and is referred to as the Keyword PDF. A corpus of such Keyword PDF files make up the test set. The generated Classification model (Classifier) is tested using the test set i.e. A classification label -the Domain/Sub domain is predicted for each of the Keywords PDF file(s).

The steps involved for Testing the Classifier can be stated as under:

- Extract Title and Keywords from each Research document and store in a separate PDF file - Keyword PDF(s) i.e. keyword pdf_1= $title_1$ + $keywords_1$
- Build a corpus of PDF(s) i.e. Cr = { keyword pdf_1 + keyword pdf_2 … keyword pdf_n }
- Preprocess Cr For each Keyword PDF in C
  - a) Remove stop words
  - b) Apply Stemming Algorithm
  - c) Generate VSM or a TD matrix - D(i, j) (where i is the Keyword PDF and j is the jth term of Keyword PDF i.)
- Apply the classification model to generate class labels for each Keyword PDF.
- Store the above results of Classification test in a temporary Data base table (keyword_temp).

(c) Process for Storing Predictions (Results) and Generating key

The results generated by process b are captured in a temporary Database table. This temporary data is further processed and a classification key is generated based on the Classification label predicted in process b. The final database table is then created with the updated data. A Java program using the JDBC API is proposed to access the temporary Database table.
The algorithm of this process can be stated as under:

Algorithm:

Reading the temporary table, generating the classification key and storing results in the final DB table.

Input:

The temporary DB table (keyword_temp - generated with Classification experiment in Rapid Miner))

Output: The final Keyword table

Steps:

```
For each record in the temporary DB table
Read keywords from the associated Keyword PDF;
Read predicted class label from temporary DB table -   keyword_temp; Generate class key using the predicted class label;
Insert a new DB record in the final Keyword_table;
end;
```

Experiment of evel-2

1. DSL files were created for all the possible domains/ sub domains.
2. PDF files were constructed with the Title and the keywords in the Research paper. Rapid Miner was used to carry out this classification experiment.
3. The K-NN classifier was trained with the help of a Corpus comprising of 24 such DSL files. Each DSL file was assigned a label which indicated its domain/ sub domain.
4. The DSL Corpus was processed to produce Term Document matrix using Term Frequency. The processing involved a) Tokenizing   b) Removing Stop Words and c) Stemming.
5. The classifier was then tested using PDF files to arrive at the predictions.
6. The Keyword table is generated at the end of this process.

Conclusion

This research paper "Text Mining Techniques: A semantic Approach in order to classify the documents "demonstrates an effective and efficient technique of classifying Text documents. The entire contents of the document need not processed to derive classification labels. This research shows how classification can be achieved by handling limited but relevant portions of the document. Classification does not necessary require a huge set of labeled training examples. A semantic classification can be achieved with a small set of DSL files and a small Concept Matrix. Though the paper suggests techniques for classifying Research papers pertaining to Computer Science; the application of this concept is not limited to such papers. It can be easily extended to other semi structured text documents. The proposed Classification framework assigns only a single classification label to each research paper. However, a single research paper may relate to 2 or more areas or domains and may contain multiple concepts. Assigning multiple labels or performing Fuzzy classification is an aspect that can be explored further.

References

[1]. Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.

[2]. Sergio Bolasco, Alessio Canzonetti, Francesca Della Ratta-Rinald and Bhupesh K. Singh, (2002), "Understanding Text Mining:a Pragmatic Approach", Roam, Italy.

[3]. Haralampos Karanikas and Babis Theodoulidis Manchester, (2001), "Knowledge Discovery in Text and Text Mining Software", Centre for Research in Information Management, UK.

[4]. Jiawei Han and Micheline Kamber "Data Mining - Concepts and Techniques" 2^{nd} Edition Thomas.

[5]. W Miller "Data and Text Mining - A Business Applications Approach" Mike Chapple, "About.com" Guide.

[6]. Tan, Steinbach, Kumar "Introduction to Data Mining" Stat Soft Electronic Statistic Text Book.

[7]. R. Feldman, "Practical Text Mining", PKDD-98, p487, 1998

[8]. Aurangzeb khan, Baharum Baharudin, Khairullah khan "Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification" 2010 Second International Conference on Computer Engineering and Applications

[9]. A. Wilcox, G. Hripcsak, and C. Friedman, "Using Knowledge Sources to Improve Classification of Medical Text Reports", (poster) KDD-2000 Workshop on Text Mining, 2000.

[10]. A. Dasgupta, "Feature selection methods for text classification." In Proceedings of the 13th ACMSIGKDD international conference on Knowledge discovery and data mining, pp. 230 -239, 2007.