Discriminative Features Selection in Text Mining Using TF-IDF Scheme

Ms. Vaishali Bhujade[1], Prof. N. J. Janwe[2], Ms. Chhaya Meshram[3]

[1]B.D.C.O.E. Sevagram, [2]R.G.C.E.R.T. Chandrapur, [3]B.D.C.O.E. Sevagram

Abstract- This paper describes technique for discriminative features selection in Text mining. 'Text mining' is the discovery of new, previously unknown information, by computer. Discriminative features are the most important keywords or terms inside document collection which describe the informative news included in the document collection. Generated keyword set are used to discover Association Rules amongst keywords labeling the document. For feature extraction Information Retrieval Scheme i.e. TF-IDF is used. This system uses previous work, which contains Text Preprocessing Phases (filtration and stemming). This work serves as basis for Association Rule Mining Phase. Association rule mining represents a Text Mining technique and its goal is to find interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored in databases, many companies are becoming interested in mining association rules from their databases to increase their profits Knowledge discovery in databases (KDD) is the process of finding useful information and pattern in data.

***Keywords*-Data Mining, Text Mining, Knowledge Data Discovery, Association Rules.**

I. INTRODUCTION

The abundance of text data generates the appearance of a new field named text mining. Text collected in large databases becomes raw material for these knowledge discovery techniques and mining tools for "gold" were necessary. The current expert system technologies, which typically rely on users or domain experts to manually, input knowledge into knowledge bases. This procedure contains errors, and it is extremely time-consuming and costly. Text mining tools which perform text analysis may uncover important text patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research [1, 2, 4]. Text mining represents the automatic process to discover patterns and relations between text data stored in large databases called warehouses, the final product of this process being the knowledge, meaning the significant information provided by the unknown elements [6].

One of the most popular text mining techniques is association rule mining. The patterns discovered with this text mining technique can be represented in the form of *association rules* [5, 4]. *Rule support, confidence, lift* and *conviction* are two measures of rule interestingness. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts.

II. TEXT MINING SYSTEM ARCHITECTURE

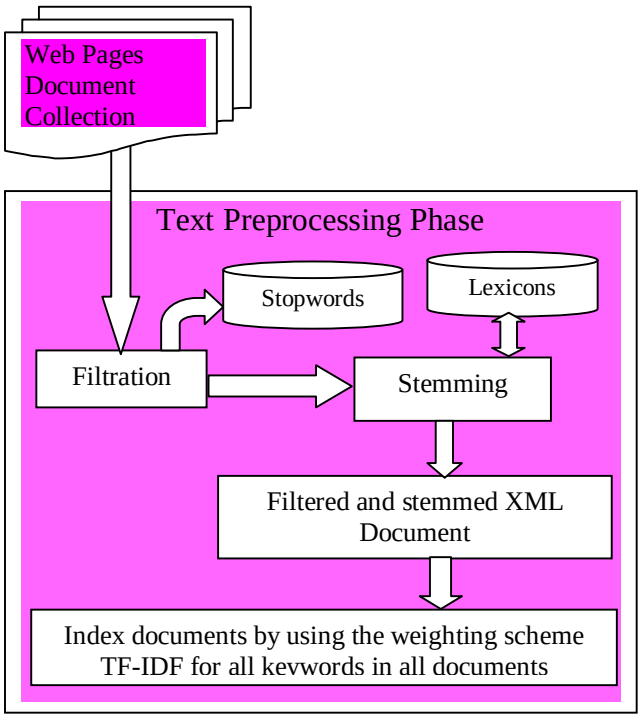

Fig 1:- Text Mining System Architecture

The text mining system, Extracting Association Rules from Text (EART) is shown in fig 1. It automatically discovers association rules from textual documents. This Text Mining system is totally depends on Information Retrieval Scheme (TF-IDF). The EART system ignores the order in which the word occurs, but instead focusing on the words and their statistical distributions. The system begins with selecting collections of documents from the web. The EART system consists of three phases: Text Preprocessing Phase (Filtration, Stemming and Indexing of the documents), Association

Rule Mining (ARM) Phase, and Visualization Phase (Visualization of results). For filtration we have different

filters(case conversion filter, stop word filter, stop symbol filter, white space filter, HTML tag filter, script tag filter) [8]. And for stemming system uses JWNL Library.

III. INDEXING

The filtered and stemmed documents are then index by using the weighting scheme. If the textual data is indexed, either manually or automatically, the indexing structures can be used as a basis for the actual knowledge discovery process. As a manual indexing is a time-consuming task [14, 15], it is not realistic to assume that such a processing could systematically be performed in the general case. Automated indexing of the textual document base has to be considered in order to allow the use of association extraction techniques on a large scale. Techniques for automated production of indexes associated with documents can be borrowed from the Information Retrieval field [13]. Each document is described by a set of representative keywords called index terms. An index term is simply a word whose semantics helps in remembering the document's main themes [13]. It is obvious that different index terms have varying relevance when used to describe document contents in a particular document collection. This effect is captured through the assignment of numerical weights to each index term of a document.

The techniques for automated production of indexes associated with documents usually rely on frequency-based weighting schemes. The weighting scheme TF-IDF (Term Frequency, Inverse Document Frequency) is used to assign higher weights to distinguished terms in a document, and it is the most widely used weighting scheme which is defined as (cf. [10] [9] [11]):

$$w(i,j) = tfidf(d_i t_j) = Nd_i t_j * log_2 (|C| / Ntj)$$

where, Nd_i,t_j denotes Term Frequency- the number of times a term occurs in a document is called its 'term frequency'. It denotes the number the term t_i occurs in document d_j. And Ntj denotes Inverse Document Frequency - the number of documents in collection where the considered term occurs at least once. And |C||- denotes the total no. of documents in Collection.

Our aim is to identify and filter the keywords that may not be of interest in the context of the whole document collection either because they do not occur frequently enough or they occur in a constant distribution among the different documents. Our system uses a statistical relevance-weighting function that assigns a weight to each keyword based on their occurrence patterns in the collection of documents, and the top *N t*aken as the final set of keywords to be used in the ARM phase.

IV. TF-IDF (Term Frequency – Inverse Document Frequency)

A. Term Frequency

The number of times a term occurs in a document is called its '*term frequency'. The Term Frequency is defined as:*

$$TF = Nt_i, d_j$$

Where Nt_i,d_j *- denotes the number the term* t_i *occurs in document* d_j

B. Inverse Document Frequency

The number of documents in collection where the considered term occurs at least once. The Inverse Document Frequency is defined as:

$$IDF = log_2 (|C|/Nt_j)$$

Where Nt_j*- denotes the number of documents in collection C in which tj occurs at least once &* $|C|$*- denotes the total no. of documents in Collection*

$$(TF\text{-}IDF)_{ij} = (TF)_{ij} * (IDF)_{ij}$$

E. g.
Consider a document containing 100 words wherein the word "*cow*" appears 3 times.
The term frequency (TF) for *cow* is:

$$TF = (3 / 100) = 0.03$$

Now, assume we have 10 million documents and *cow* appears in one thousand of these. Then, the inverse document frequency is calculated as

$$IDF = \log (10\,000\,000 / 1\,000) = 4$$

The TF-IDF score is the product of these quantities:

$$TF\text{-}IDF = 0.03 \times 4 = 0.12.$$

V. EXAMPLES

We have extracted 5 web documents from the internet. For this connectivity we use Yahoo API. We extract web documents related to Cryptography. The experiment result has shown below:

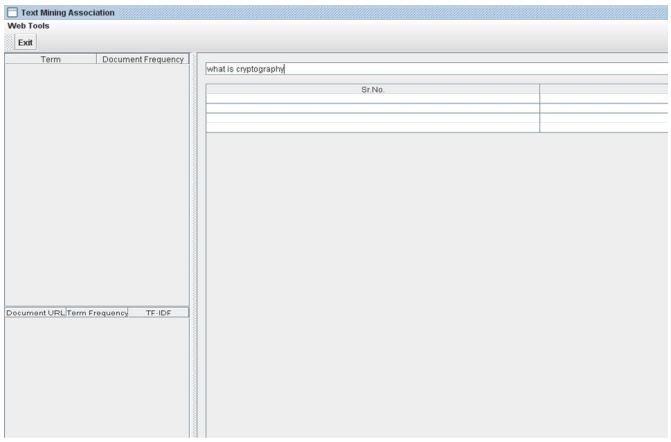

Fig 1: Extraction of Data from Search Engine

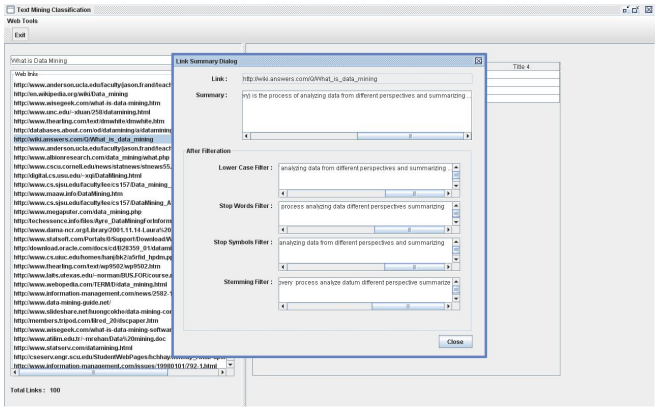

Fig 2: Text Preprocessing Phase (i.e. Filtration & Stemming)

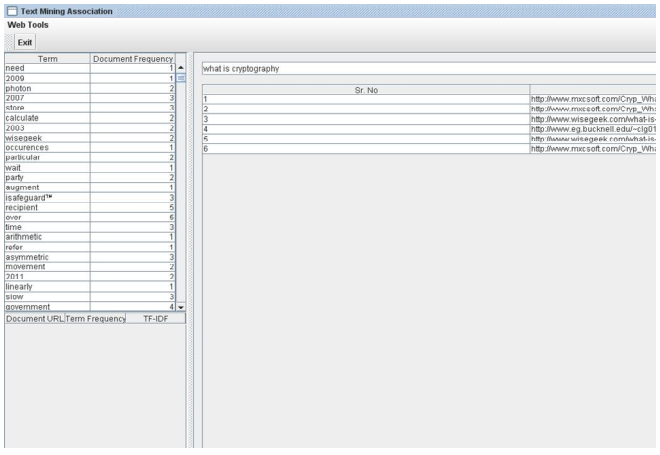

Fig 3: Indexing (using TF-IDF scheme)

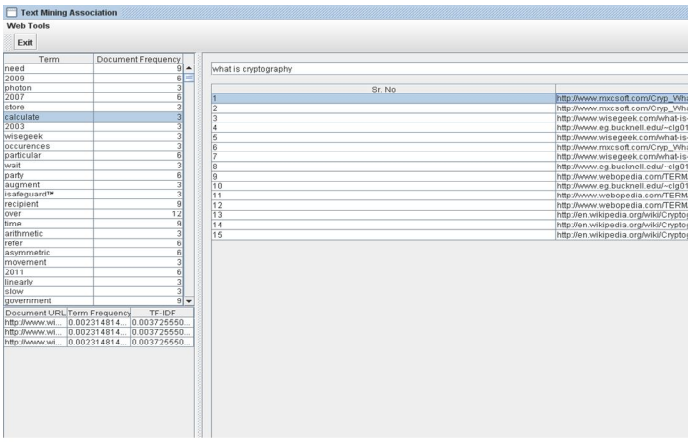

Fig 3: Indexing (using TF-IDF scheme)

VI. CONCLUSION

This text has introduced a new branch of data mining - text mining. It has mainly discussed some key points about discriminative features selection in text mining. The whole process is based on TF-IDF scheme. This paper has presented a technique for TF-IDF of keywords. Technique based on keyword features. The EART system ignores the order in which the word occurs, but instead focusing on the words and their statistical distributions. This paper also presents a new technique for filtration of unimportant data from collection of documents, And Stemming of documents. For filtration system uses number of different filters. In this paper we showed that general Information Retrieval Scheme methods are applicable to text analysis tasks. Moreover, we presented a general framework for text mining. The framework follows the general KDD process, thus containing steps from preprocessing to the utilization of the results. We gave snapshots of examples of how to do preprocessing.

VI. REFERENCES

[1]. A g r a w a l , R . S r i k a n t , R . - Fast Algorithms for Mining Association Rules, *Proc. of the 20th Int'l Conference on Very Large Databases*, Santiago, Chile, 1994

[2]. Fa y y a d, U. M., P i a t e t s k y - Sh a p i r o , G . , Smy t h , P . , Ut h u r u s a m y, R . -*Advances in Knowledge Discovery and Data Mining*, AAAI Press Series in Computer Science. A Bradford Book, the MIT Press, Cambridge Massachusetts, London Englan, 1996

[3]. Fa y y a d, W., P i a t e t s k y - S h a p i r o, G., S m y t h , P . - From data mining to knowledge discovery: An overview, In: *Advances in Knowledge Discovery and Data Mining*, W. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), AAAI/MIT Press,Cambridge/USA, pp. 1 – 3, 1996

[4]. H a n, J., F u, Y. – Discovery of Multiple-Level Association Rules from Large Databases, *Proc.of 1995 Int'l Conf. on Very*

Large Data Bases (VLDB'95), Zürich, Switzerland, September 1995, pp.420-431, 1995
[5]. S r i k a n t , R . , A g r a w a l , R . - Mining Generalized Association Rules, *Future Generation Computer Systems*, 13(2-3), 1997
[6]. * * * - *Data Mining*, CINECA site, http://open.cineca.it/datamining/, accessed 15.01.2008
[7] H. Mahgoub,"Mining association rules from unstructured documents" in *Proc. 3rd Int. Conf. on Knowledge Mining, ICKM*, Prague, Czech Republic, Aug. 25-27, 2006, pp. 167-172.
[8] Ms. Vaishali G. Bhujade, and Prof. N. Janwe, "OBSOLESCENCE DATA REMOVAL IN TEXT MINING," International ConferenceICISET,8-9April 2011
[9] J. Paralic and P. Bednar, "*Text mining for documents annotation and ontology support* (*A book chapter in: "intelligent systems at service of Mankind," ISBN 3-935798-25-3, Ubooks*, Germany, 2003*)*.
[10] C. Manning and H Schütze, Foundations of statistical natural language processing (MIT Press, Cambridge, MA, 1999).
[11] M. Rajman and R. Besancon, "Text mining: natural language
techniques and text mining applications", in *Proc. 7th working conf.*
on database semantics (DS-7), Chapan &Hall IFIP Proc. Series.
Leysin, Switzerland Oct. 1997, 7-10.
[12] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo,editors, Proc. 20th Int. conf. of very Large Data Bases, VLDB,*Santigo, Chile, 1994, 487-499.
[13] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval(Addison-Wesley, Longman publishing company, 1999).*
[14] R. Feldman and I. Dagan, "Knowledge discovery in textual databases (KDT)", in *Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining, 1995.*
[15] R. Feldman and H. Hirsh, "Mining associations in text in the presence of background knowledge," in *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, Portland, USA, 1996.