ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Discovering Relations among Documents Using Novel Text Retrieval Technique

Manjiri Gajanan Ghadi[1], Carmen Lysandra Pereira [2], Manimozhi R. [3]

Department of Software Technology, AIMIT, St Aloysius College, Mangalore, India[1]

Department of Software Technology, AIMIT, St Aloysius College, Mangalore, India[2]

Associate Professor, Department of MCA, AIMIT, St Aloysius College, Mangalore, India[3]

ABSTRACT: Text categorization is one of the key techniques in text mining to categorize the documents in a supervised manner.In this paper we have done study on automaticcategorization of news items.The categorization algorithm transforms each document into a vector of weights corresponding to an automatically extracted set of keywords. This process is performed on a large set of news items, forming the multi-dimensional space populated by news items of known categories. An unknown news item is also transformed into a vector of keyword weights and then categorized using the k-means method in this space. Finally the documents are compared based on weighted keywords to find which documents are most similar.

I. INTRODUCTION

Over the last decade, the emergence of the world wide web has led to an exponential increase in amount of digital documents available for various purposes has grown enormously with the increasing availability of high capacity storage hardware and powerful computing platforms. The vivid increase of documents demands effectual organizing and retrieval methods mainly for large documents. This development has led to an increase interest in methods that allows the users to quickly and accurately retrieve and organize these types of information.

Automatic text categorization has always been an important application and research topic since the inception of digital documents. Today, text categorization is a necessity due to the very large amount of text documents that we have to deal with daily(Maria-Luiza Antonie University of Alberta). The applications range from automatic document indexing for information retrieval systems, document organization, text filtering, word sense disambiguation, categorization of web pages and most recently spam filtering.

In general, automated text categorization may be regarded as two separate tasks. First, document indexing is neededin order to transform the natural language text into a numerical representation suitable for further processing. The second task is the actual categorization.

Document indexing consists of choosing the appropriate set of keywords (feature selection) based on the wholecorpus of documents, and assigning weights to those keywords for each particular document, thus transforming each document into a vector of keyword weights. Section 2 describes the process of feature selection and weight calculation for document indexing.

II. DOCUMENT INDEXING

The categorization algorithm transforms each document into a vector of weights that are equivalent to an automatically chosen set of keywords.This consists of two main steps.First the appropriate set of keywords has to be chosen .this set of keywords is used for all the documents to be indexed and therefore it has to be universal, i.e. cover all necessary words that can be characterize the documents.This step is performed only once the second step is performed on each document to be indexed and consists of assigning a weight to each of the keyword such that the weight determines the relative frequency of occurrence of the keyword in the indexed document.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

2.1ELECTING KEYWORDS

Keyword s provide a concise and precise high level summarization of a document .keywords constitute an important feature for document retrieval, classification, topic search and other tasks.Electing keywords is the main preprocessing step necessary for document indexing. It is essential in determining the quality of later classification. It is important to select the significant keywords that carry the meaning and discard the words that donot contribute to distinguishing between documents.

Electing keywords process is based on scrutinizing the whole corpus of available documents. It discards both the word that appears infrequently in the corpus and the words that appear very often.If a word appears only a very few times in the whole corpus it is likely to be a spelling error or some special case that will not contribute to the quality of classification.On the other hand words that appear frequently in the corpus across all categories are likely to be too general to be used for classification.The process runs as follows:

First all document in the corpus are processed and the matrix of all words and their appearance count is obtained. Punctuations, numbers and special characters are discarded. At the end of the process we obtain the set of characteristic keyword that can be used to assume category of a given document. To reduce the dimensionality of the set of documents word stemming method are used.

2.2FINDING WORD STEM

Information Retrieval (IR) is essentially a matter of deciding which documents in a collection should be retrieved to satisfy a user's need for information. The user's information need is represented by a query or profile, and contains one or more search terms, plus perhaps some additional information such importance. Hence, the retrieval decision is made by comparing the terms of the query with the index terms (important words or phrases) appearing in the document itself. The decision may be binary (retrieve/reject), or it may involve estimating the degree of relevance that the document has to the query.

Unfortunately, the words that appear in documents and in queries often have many morphological variants. Thus, pairs of terms such as "computing" and "computation" will not be recognized as equivalent without some form of natural language processing (NLP).

In most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. For this reason, a number of so-called stemming Algorithms, or stemmers, have been developed, which attempt to reduce a word to its *stem* or root form. Thus, the key terms of a query or document are represented by stems rather than by the original words. This not only means that different variants of a term can be conflated to a single representative form – it also reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time.

Example the words playing, played can be stemmed to the word play.
In the present work porter stemmer algorithm is used which is most commonly used algorithm.

2.2.1 PORTER STEMMER

Porter Stemmer is based on the idea that the suffixes in the English language (approximately 1200) are mostly made up of a combination of smaller and simpler suffixes. This Stemmer is a linear step Stemmer. Specifically it has five steps applying rules within each step. Within each step, if a suffix rule matched to a word, then the conditions attached to that rule are tested on what would be the resulting stem, if that suffix was removed, in the way defined by the rule.Once a Rule passes its conditions and is accepted the rule fires and the suffix is removed and control moves to the next step. If the rule is not accepted then the next rule in the step is tested, until either a rule from that step fires and control passes to the next step or there are no more rules in that step whence control moves to the next step. This process continues for all five steps, the resultant stem being returned by the Stemmer after control has been passed from step five. See figure 1 (Rob Hooper and Chris Paice, 2005).

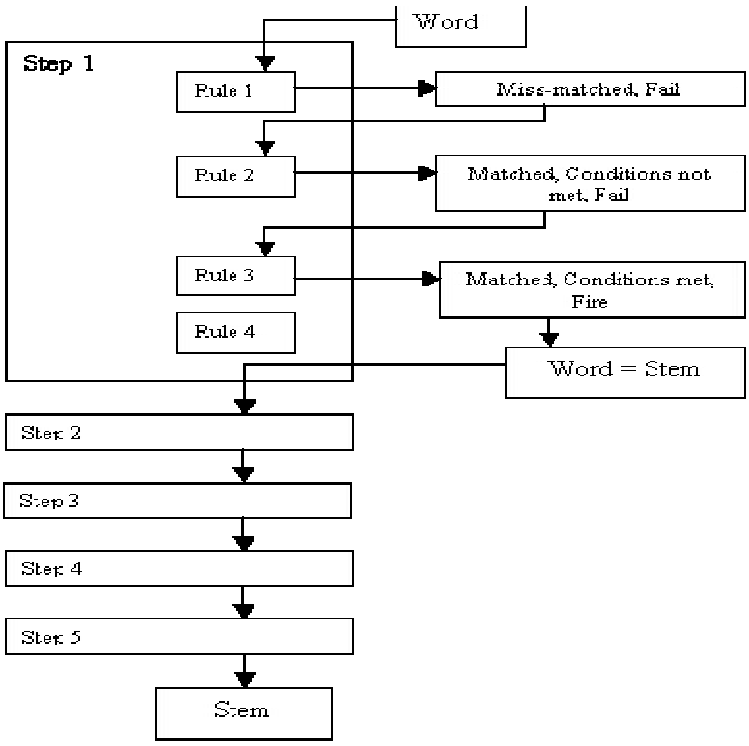
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Figure 1 Overview of Porter algorithm



2.3 ASSIGNING WEIGHTS TO KEYWORDS

In order to represent a document by a numeric vector we need a function that assigns a weight factor to each of the keywords chosen in the preprocessing steps.th e weight factors should represent the importance of the keyword for the classification of the document.

In the application td-idf (term frequency-inverse document frequency is used. Tf-idf is a numerical statistic which reveals that a word is how important to a document.Tf-idf is often used as a weighting factor in information retrieval text mining.

The value of tf-idf increases propositionally to the number of times a word appears in the document but is counteracting by the frequency of the keyword in the corpus. This can help to control the fact that some words are generally more common than others. Tf-idf can be successfully use for step words filtering in various subject field, text summarization and classification

Tf-idf is the product of two statics which are term frequency and inverse document frequency. To further distinguish them the number of times each term occurs in each document is counted and sums them altogether.
Term frequency is defined as number of times a term occurs in a document
Term represents the keyword.

If (t, d) =0.5+ (0.5*f (t, d)/maximum occurrence of words) *(Equation I)*

Inverse document frequency, it is a statistical weight used for measuring the importance of a term in a text document collection. Idf feature is Inco-operated which reduces the weight of terms that occurs very frequently in the document set and increases the weight of terms that occur rarely.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

Idf(t, d) =log (|D|/number of documents term t appears) (*Equation II*)

Then term frequency inverse document frequency is calculated for each word using the formula.

Tf-idf (t, d, D) =tf (t, d)*idf (t, d, D) *(Equation III)*

In this equation I and II $f_{t.d}$denotes the frequency of the occurrence of term t in document d. in equation III tf-idf is calculated for each term in the document by using term frequency (tf_{td}) and inverse document frequency (idf_{td}). (N, Text Classification using Keyword Extraction Technique, December 2013)

2.4SIMILARITY MEASURING AMONG THE DOCUMENTS

In similarity measuring we are comparing the documents based on the assigned weights of the keywords which is described in section 2.1 to find the relevance among the documents which are most similar.

2.4.1 CLUSTERING

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression.

Cluster analysis is an important human activity. Early in childhood, we learn how to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes. By automated clustering, we can identify dense and sparse regions in object space and, therefore, discover overall distribution patterns and interesting correlations among data attributes. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns.

Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for outlier detection, where outliers (values that are “far away” from any cluster) may be more interesting than common cases.

2.4.1.1 K-MEANS ALGORITHM

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. K-Mean clustering solves $J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$,

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j ,is an indicator of the distance of the *n* data points from their respective cluster centers.

The algorithm is composed of the following steps:

1. *Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.*
2. *Assign each object to the group that has the closest centroid.*
3. *When all objects have been assigned, recalculate the positions of the K centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

The below figure 2 shows the overview of text categorization

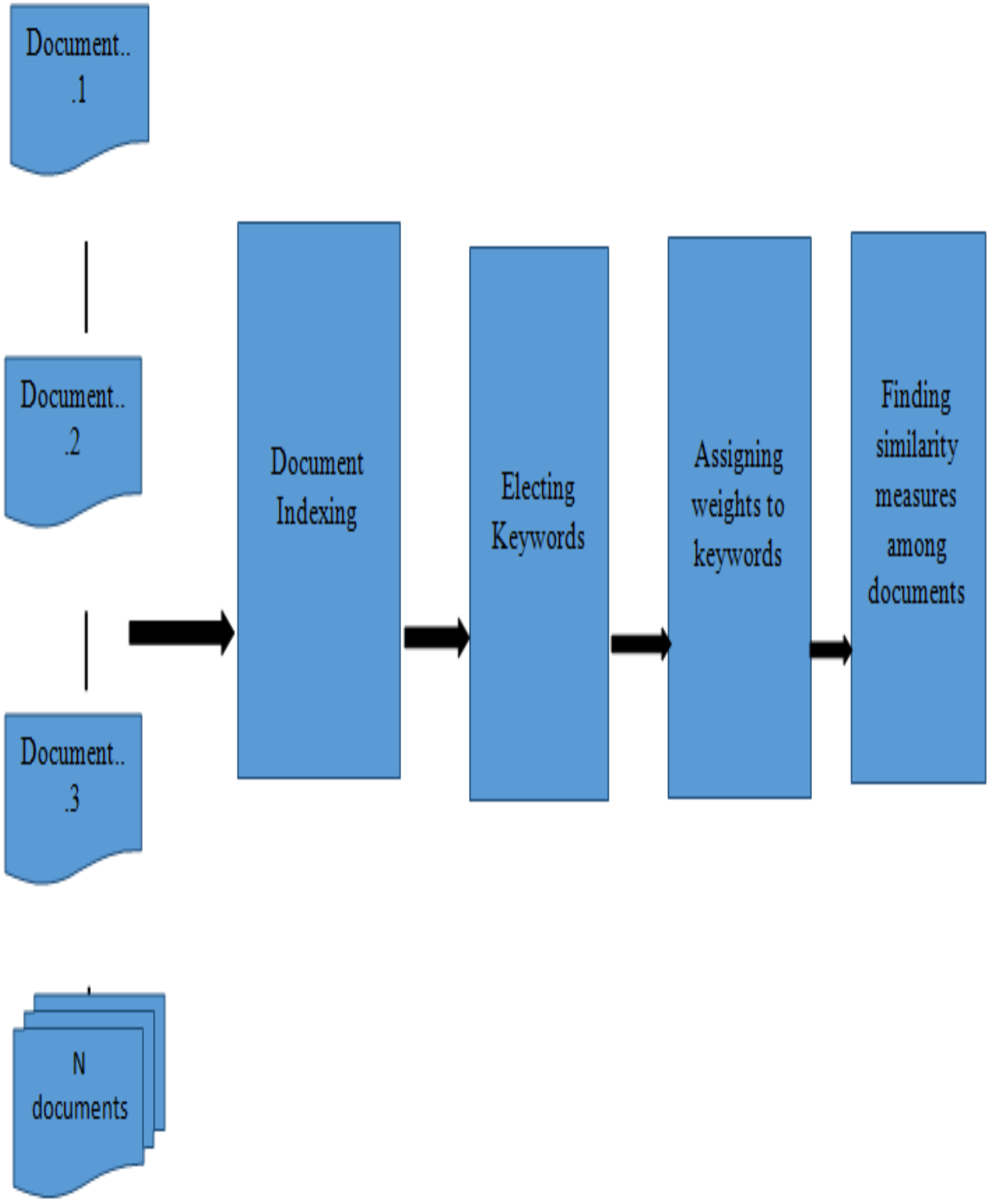


Figure 2.Overview of Automatic Text Categorization

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

III. TEXT CATEGORIZATIONIN NEWS INDUSTRY

There are great amounts of textual information on the internet and in computerized systems ofdifferent associations and companies, it is very expensive to manually manipulate this huge amount of information for that it is very time consuming and needs expensive expertise who cannot be available for 24 hours a day, automatic text categorization helps in reducing the time needed to classify hundreds or thousands of daily arrived documents, without the need for experts (Dr. Riyad Al-Shalabi).Since the development in technology has decreased the complexity of news exchange and newspaper printing, it has resulted in dramatic increase in the number of available news sources and the volume of news items an average recipient are receiving every day. In the news industry metadata is a very part of a news items.

On the other hand, speed has always been very important factor in the new industry. Due to the inability to process all the content they receive fast enough, news recipients have to rely on metadata to find out the content they are interested in, which means that it is very important for metadata to be consistent, accurate and comprehensive.

The different categories in the news industry are:

DEFINED CATEGORY	CATEGORY CONTENT
Category 1	Business News
Category 2	National News
Category 3	International News
Category 4	Sports News
Category 5	Technology News

IV. IMPLEMENTATION

The system is implemented using the RapidMiner tool.We have used sports related documents to find the similarity among the documents. All the documents are uploaded in the RapidMiner tool. Then we are performing all the text preprocessing steps whichincludes tokenization,stopword,transform cases, and stemming and then we are finding the similarity among the documents using K-Mean clustering algorithm.

V. RESULT

The system was tested on randomly chosen news items belonging to sports category.After implementation following result was obtained.

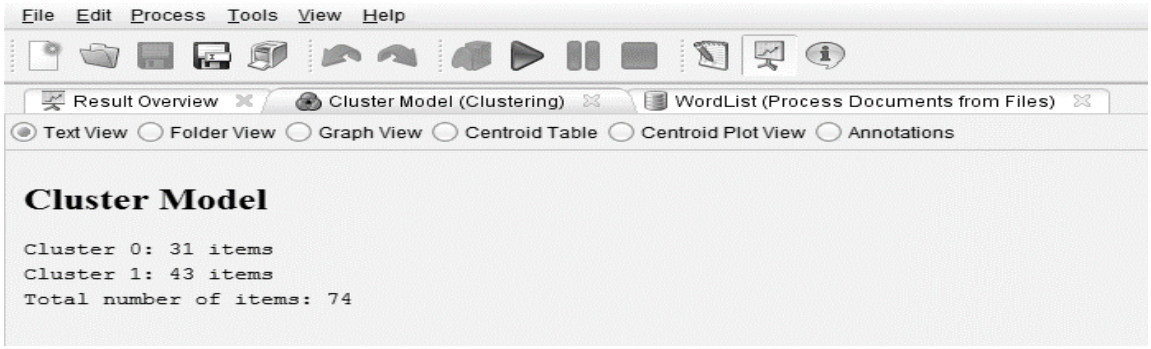


Figure 3

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

The above figure shows that there are two clusters that are obtained after the implementation of the system. There are two cluster (cluster 0 and cluster 1) each cluster contains documents that are similar to each other.

VI. CONCLUSION AND FUTURE WORK

In this paper we have presented initial results of an implementation of an automated sports categorizationsystem. The obtained results are encouraging and practically useful. Further work will continue on extending the system to cover not only the sports categories but all the categories that are available in the media, but also the sub-categories in a hierarchical categorization scheme.

VII. ACKNOWLEDGEMENT

We are indebted to many people for the successful completion of this project and would like to take this opportunity to acknowledge the effort put forth by them on our behalf.we would like to extend our hearty thanks to our guide Mrs. Manimozhi R.Asso. Prof,andMrsKavitha R, Asst. Prof. MCA Department, AIMIT and MrLanwin Lobo in succesfull completion of our research paper.

REFERENCES

1. (n.d.). Retrieved from Introduction to Text Categorization: http://users.softlab.ntua.gr/facilities/public/AD/Text%20Categorization/Introduction%20to%20Text%20Categorization.ppt
2. .:::.International Journal of Engineering Research and ... (n.d.). Retrieved from ijera: http://www.ijera.com/pages/v3no3.html
3. An Efficient Modified K-Means Algorithm To Cluster Large Data ... (n.d.). Retrieved from ijarcsee: http://www.ijarcsee.org/index.php/IJARCSEE/article/view/72
4. An Improved k-Nearest Neighbor Algorithm for Text ... (n.d.). Retrieved from citeseerx:
5. Arabic Text Categorization. (n.d.). Retrieved from ResearchGate : http://www.researchgate.net/publication/228802987_Arabic_Text_Categorization_Using_kNN_Algorithm
6. Automated News Item Categorization (2005) . (n.d.). Retrieved from citeseerx: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.102.4034
7. ch8 : CS 595 : Western Michigan : Class Note. (n.d.). Retrieved from coursehero: http://www.coursehero.com/file/1870740/ch8/
8. CiteSeerX â€" Citation Query Bommel, P ... (n.d.). Retrieved from CiteSeerX: CiteSeerX â€" Citation Query Chaitanya Kamisetty, Rajeev ... (n.d.). Retrieved from CiteSeerX :
9. Classifying (2002) - CiteSeerX. (n.d.). Retrieved from citeseerx: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.15.3778
10. DATA CLUSTRING AND TECHNIQUES (1) Technical. (n.d.). Retrieved from seminarprojects: http://seminarprojects.com/Thread-data-clustring-and-techniques
11. Data Mining | Information Cloud. (n.d.). Retrieved from itcloud: http://itcloud.net46.net/tag/data-mining/
12. Data Mining Cluster Analysis - Tutorials for COBOL, XSD ... (n.d.). Retrieved from tutorialspoint: http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm
13. Data Warehousing and Data Mining | Online Engineering. (n.d.). Retrieved from Online Engineering: http://onlineengineering.wordpress.com/category/it/data-warehousing-and-data-mining-it/
14. Dr. Riyad Al-Shalabi, D. G. (n.d.). Arabic Text Categorization. 9.
15. Hrvoje Bacan1, I. S. (2010). Automated News Item Categorization. Unska 3, HR-10 000 Zagreb, Croatia.
16. IJCA - Emotion Classification in Arabic Poetry using Machine ... (n.d.). Retrieved from ijcaonline: http://www.ijcaonline.org/archives/volume65/number16/11006-6300
17. Inverse Document Frequency - Springer Reference. (n.d.). Retrieved from springerreference: http://www.springerreference.com/docs/html/chapterdbid/63952.html
18. Kamber, J. H.-C. (Second Edition). Data Mining concepts and techniques. morgan kalifmann publishers.
19. Liu, Y. Y. (1999). A re-examination of text categorization methods. . In proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval Technique, 9.
20. Managing content with automatic document classification (2004). (n.d.). Retrieved from citeseerx: http://citeseerx.ist.psu.edu/showciting?cid=519243
21. Maria-Luiza Antonie University of Alberta, C. l. (n.d.). Text Document Categorization by Term Association. 8.
22. N, M. S. (12 December 2013). Text Classification using Keyword Extraction Technique . International Journal of Advanced Research in Computer Science and Software Engineering , 7.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol.2, Special Issue 5, October 2014

23. PPT â€“ Clustering PowerPoint presentation | free to download. (n.d.). Retrieved from powershow: http://www.powershow.com/view/17b1c5-NzFmZ/Clustering_powerpoint_ppt_presentation
24. Proceedings of the 17th annual international ACM SIGIR ... (n.d.). Retrieved from acm digital library: http://dl.acm.org/citation.cfm?id=188490
25. Rob Hooper and Chris Paice, S. 2. (2005). The Lancaster Stemming Algorithm. Retrieved from http://www.comp.lancs.ac.uk/computing/research/stemming/general/index.htm
26. Textmining Predictive Models. (n.d.). Retrieved from SlideShare: http://seminarprojects.com/Thread-data-clustring-and-techniques
27. Textmining Predictive Models. (n.d.). Retrieved from slideshare: http://www.slideshare.net/dataminingtools/textmining-predictive-models-2749248
28. tfâ€“idf - Wikipedia, the free encyclopedia. (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Tf-idf
29. tf–idf. (n.d.). Retrieved from tfâ€“idf - Wikipedia, the free encyclopedia: http://en.wikipedia.org/wiki/Tf%E2%80%93idf
30. What is Cluster Analysis? - 414945 - SlideServe. (n.d.). Retrieved from SlideServe: http://www.slideserve.com/hedya/what-is-cluster-analysis
31. What is Porter Stemming? (n.d.). Retrieved from School of Computing & Communications: http://www.comp.lancs.ac.uk/computing/research/stemming/general/porter.htm
32. What is Stemming? (n.d.). Retrieved from School of Computing & Communications: http://www.comp.lancs.ac.uk/computing/research/stemming/general/
33. .Y. Yang and X. Liu. A re-examination of text categorization methods. In proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99),42-49, 1999