Discovering Knowledge from Text Repositories Using Information Extraction: A Review

Prof. Sandeep R. Sirsat
Associate Professor and Head
Department of Computer Science
Shri Shivaji Science & Arts College, Chikhli.
Distt:- Buldana (M. S),

Dr. Vinay Chavan
Associate Professor and Head
Department of Computer Science
Porwal College, Kamptee,
Nagpur (M. S),

Dr. Shrinivas P. Deshpande
Associate Professor and Head
P. G. Department of Computer Science and Technology, DCPE,
HVPM, Amravati (M.S),

Abstract: This paper presented a review of work done by researchers for discovering knowledge from text repositories. There are two approaches to mining text form online repositories. First, when the knowledge to be discovered is expressed directly in the documents to be mined, IE alone can serve as an effective tool for such text mining. Second, when the documents contain concrete data in unstructured form rather than abstract knowledge, IE can be used to first transform the unstructured data in the document corpus into a structured database, and then use some state-of-the-art data mining algorithms/tools to identify abstract patterns in this extracted data. The main aim of this review is to discuss several methods related to these two approaches and to study the merits and demerits of them.

Keywords: Information Extraction, Text Mining, Text repositories, Knowledge Discovery from Database (KDD).

1) Introduction:-

In the contemporary era, most of the information is available in the form of unstructured natural language documents due to the growth of the web, digital libraries, technical documentation etc. It is the need of the hour to discover non-trivial, previously unknown, and potentially useful knowledge from such unstructured natural language documents. Hence discovering useful knowledge from unstructured text, i.e. text mining, is becoming an increasingly important aspect of Knowledge Discovery. [2,3,5,6,8]. Thus text mining is an increasingly important research field, and is similar in some sense with data mining, as both are included in the field of information mining. However, data mining deals with structured database. Thus it is important to transform the unstructured text into structure database, so that the data mining techniques may be adapted in a straightforward way to mine text.

Information Extraction (IE) methods, with reasonable accuracy, are able to transform the unstructured text into structured database, called intermediate forms. The most usual intermediate forms are: Bag-of-words, N-grams, Keyphrases, Multi-term phrases, Concept word, Concept Hierarchy, Conceptual graph, etc.

Most of the text mining techniques treat documents as an unordered bag-of-words. Then it specifies a weighted frequency for each of these terms in the documents as a parse vector using standard vector space model. Such a simplified representation of text has been shown to be quite effective for the number of standard tasks, such as, information retrieval (IR), classification, and clustering [15, 22].

However, mining knowledge from text cannot be discovered using simple bag-of-words representation. It is not useful to assert the properties and relationship among the important entities using standard vector-space model. Thus existing methods in IE, with reasonable accuracy, are able to identify several types of entities in text documents and establish some relationships that are asserted between them [3, 4, 5, 7, 9]. Recently developed text mining techniques describes the integration of information retrieval methods with data mining techniques for association rule discovery [2, 8].

Therefore, IE can serve as an important technology for text mining. If the knowledge to be discovered is expressed directly in the documents to be mined as in [6, 7], IE alone can serve as an effective approach to text mining. However, if the documents contain concrete data in unstructured form rather than abstract knowledge, IE can be used to first transform the unstructured data in the document corpus into a structured database, and then use some state-of-the-art data mining algorithms/tools to identify abstract patterns in this extracted data [2, 3, 5, 8].

In this paper, we reviewed these two approaches to mining text form online repositories i.e. (KDD) with information extraction. In section 2, we have discussed the basics of information extractions. Next we discuss the IE methods to directly extract knowledge from text in section 3. In section 4, we discuss the approaches/techniques of discovering knowledge by mining structured database that is first extracted from unstructured text. In section 5, we presented the future research scope in text mining using

IE and some alternative approaches. Finally we end the discussion with conclusion in section 6.

2) Information Extraction: Problems and Methods:

Information Extraction (IE) is concern with locating specific set of relevant items from natural language documents. Thus IE systems can extracts structured information from unstructured text. One type of IE is *named entity extraction* and then creation of filled templates [18, pp 157-157]. The named entity extractor indentifies references to particular kinds of objects such as names of peoples, companies, and locations. Huanzhong Duan and et.al studies the features of the Chinese Named Entity Recognition (CNER) based on conditional Random Fields (CRFs) [16]. These features include common attributes, feature templates varying in windows size (3, 5, 7) and sequence labels sets. In addition to recognize entities, it is useful to specify specific types of relations between entities. For example, KnowITALL system is able to extract instances of relations, such as capitalOF(City, Country), or starsIN(Actor, Film) [20]. Doug Downey and et.al apply a simple pattern learning algorithm, to the task of information extraction that can be used as both extractors (to extract the instances of relations) and discriminators (to access the truth of extracted information) [11]. J. Callan an et.al. present a new approach to named-entity detection, known as KENE, which uses knowledge-based approach for learning extraction rules[16]. It uses generate-and-test approach to named-entity extraction from structured documents. *Andrew Carlson and et.al*, consider the problem of semi-supervised learning to extract categories (e.g. academic fields, athletes) and relations e.g. PlaySport(athlete, sport) from web pages, starting with a handful labeled seed example for each, and a set of constraints that couple the various categories and relations[7]. This approach shows that, training both contextual pattern extractor that extract information from freeform text (e.g. the pattern "Mayor of arg1" as an extractor for the category <city>) and wrapper which extracts information from semi-structured documents (e.g. the wrapper "<td> Class = "City" > arg1</td>" from some specific URL).

It can also be used to extract fillers for a predefined set of slots (rules) in a particular template (frame) relevant to the domain. RAPIER considers the task of extracting database from posting to the USENET newsgroup, *austin.Job* [6]. This system uses an automatic pattern based learning approach that extracts rules for indentifying each type of entity or relation. The learned extraction rules consist of three parts. 1) a pre-filler pattern that matches the text immediately preceding the phrase to be extracted. 2) a filler pattern that matches the phrase to be extracted, and 3) a post filler pattern that must match the text immediately following the filler. It uses regular expression (regexs) languages similar to that used in Perl [19], to express the pattern that utilizes limited syntactic information, produced by POS tagger [18]. It consists of a specific to general (bottom-up) search for pattern that characterizes slot fillers and their surrounding context.

Another approach of IE is to extract structured data (pattern) from unstructured or semi-structured web pages. Pattern induction to generate extraction patterns from a number of training instances is one of the most widely applied approaches to IE. *Seokhwon in et.al* proposed a local tree alignment based soft pattern matching method for IE [13]. This method considers the node labels, as well as link labels to the head node, because the class of link to the head node is important as the node label itself for dependency trees.

Moreover, the method also considers the alignment of slot value nodes in the tree patterns for adapting information extraction task. If the pattern node v is a kind of slot value nodes, the similarity score between v and w is inherited from parents of both nodes. It then constructs the pattern candidate sets for four types of pattern representation models, based on the dependency trees and scenario templates of the training data. For each pattern candidate, corresponding confidence score and optional threshold value were computed and arranged them in descending order of confidence score.

An another general approach to IE is to treat it as a sequence labeling task in which each word (token) in the document is assigned a label (tag) from a fixed set of alternatives one approach to the resulting sequence labeling problem is to use a statistical sequence model, such as hidden Markova model (HMM) [18], or a conditional Random field (CFR) [17].

3) **Extracting Knowledge From Text:**

If the information extracted from a corpus of documents represents abstract knowledge rather than concrete data, IE itself can serve as 'discovering' knowledge from text. In this section, we review some approaches/methods which discover knowledge by extracting information, such as, keyphrases/keywords extraction from text. These would be useful for many other text mining tasks, such as, classification, clustering, summarization, topic detection, etc.

One of the approaches to extract relevant information from the related topic is the selection of the one or more phrases that best represent the content of the documents. Unlike the IE task, it does not consider any known fields (slots) for a template. Instead, text segments that are unique and most representative of documents are extracted. Most systems use TF-IDF scores to sort the phrases in multiple text documents and select the top-k as keyphrases. Many existing methods convert the keyphrase extraction as classification problem using S.V.M. [15, 22]. *Zhuoye Ding and et.al* proposed a novel formulation, which present several criteria of high quality news keyphrase and integrate those criteria into the keyphrase extraction task by converting the task to the binary integer programming (BIP) problem [6]. In this approach the proposed BIP based method can combine the unsupervised methods, such as, TF-IDE and locality information, as assignment value in the object function. The locality information represents the occurrence position of the candidate phrase, either in the title of news article or in the first sentence. Further, to extract quality set of phrases, this method considers several constraints converted from the coverage and coherence criteria, and the number of extracted phrases. It assumes that high quality keyphrases should cover the whole document or group of documents in a right order. First, to satisfy coverage criteria, the Latent Dirichlet Allocation (LDA) model is used to estimates words distribution over topic. Second, it uses mutual information (MI) to measure the word coherence, which should satisfy other criteria that the keyphrases should be semantically related and coherent. In this case, the keyphrases pairs with high occurrence frequency are selected together. An experimental result proves that TF-IDF is the most important feature and locality feature can further improve the performance.

Another approach presented by I. Bhattacharya and et.al, focuses on dictionary-based text mining and its role in enabling practitioners in understanding and analyzing large text datasets [12]. To build a concept dictionaries for annotating a collection of documents form a particular domain, they define dictionary D= *Dict*(C, X) as a set of words that refers to or describes a semantic concept C in a document collection X. In this method, an online interactive framework is applied, where the user starts off with a small set of words, inspects the results, selects and rejects words from the returned ranking, and iterates until get satisfied. With interactive supervision, the user provides positive and negative seed words at each stage of iteration to the algorithm. This process gradually refines the seed sets and the ranking comes closer to the user's preference as the iteration continues.

This framework of constructing dictionary needs to provide a set of seed words for specifying a concept C and refers the WordNet to define the semantics of seed set unambiguously, for general purpose English words. However, ambiguities may arise in selecting the words in seed set, or some subset of them, when all words are not identical to these concepts or the conceptual structure is absent like WordNet. Further, the dictionaries need to be constructed for every new dataset and the existing concept nodes can be used for seeding. Then the ranking returned by the system is inspected to create the adapted dictionary for the new document collection. Thus re-using dictionaries can significantly make easier the task of specifying the semantic concept in the absence of semantic structure for dictionary construction for a concept.

The observation made from the experimental results suggest that–Interactively building dictionaries from scratch leads to good dictionaries, but adapting earlier dictionaries also leads to dictionaries of similar quality and adapting dictionaries consistently results in 50-60% time savings. One shortcomings of this supervise model is its need to provide the good quality positive and negative set of seed words. Again the system is unable to measure benefits in terms of precision and recall, as extensive experimentation is required due to lack of public tagged corpora.

4) Discovering Knowledge From Text:

In many cases, the information extracted from unstructured text represents specific data rather than abstract knowledge. In such situation, the text mining task requires to perform some additional process to mine knowledge from this specific data. [3, 5]. One approach to text mining is to first use IE to obtain structured data from unstructured text and then use traditional KDD tools to discover knowledge from this data. R. J. Mooney presents a framework for text mining, called DiscoTEX (Discovery form Text EXtraction) [3] as shown in figure 1. It uses learned information extraction system to transform unstructured text into more structured data and then mine this data to form interesting relationship.

The IE learning system of DISCOTEX integrates IE module acquired by an IE learning system, and a standard rule induction module. The IE module used two state-of-the-art system for learning information extractors, RAPIER (Robust Automated Production of Information Extraction Rules) [4] and BWI (Boosted Wrapper Induction). In section 2, we reviewed RAPIER system which performed well on realistic applications such as USENET job posting. After constructing an IE system that extracts the desired set of slots for a given application, IE extraction patterns is applied to each document of a text corpus to create a collection of structured records (database). Standard KDD techniques can then be applied to the resulting database to discover interesting relationship. Specially, DISCOTEX induce rules for predicting each piece of information in each database field given all other information in a record. To discover prediction rule, each slot value pair in the extracted database is treated as a distinct binary feature.

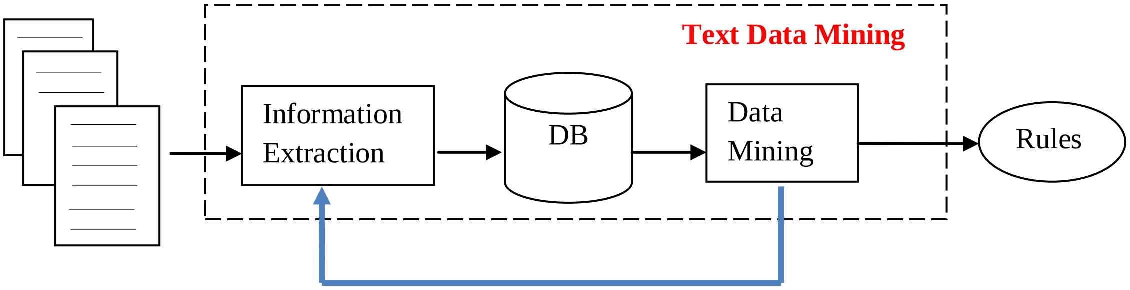

Figure 1: Overview of IE-based text mining framework

It then applies C4.5 RULES to discover interesting rules from the resulting binary data. Discovered knowledge describing the relationship between slot values is written in the form of production rules. For example, If there is a tendency for “Web” to appear in the ‘area’ slot, when “Director” appears in the ‘application’ slot, this is represented by the production rule.

“Director ∈ application ⟶ Web ∈ area”

The major drawback of DISCOTEX is that, it focuses on rules predicting the presence of fillers rather than predicting the absence of filler in a slot.

Chris Clifton, and et.al. developed an another technique TopCat (Topic Categories) for identifying topics that recur in articles of text corpus [5]. This method used IE to identify named entities in individual articles and represent them as a set of items of an article. Thus they view the problem in data mining/database context, by identifying frequent itemsets that is group of named entities that commonly occurred together. TopCat use association rule data mining technique for identifying these frequent itemset. It further clusters the named entities, using a hypergraph splitting technique, which finds, group of frequent itemsets with considerable overlap. Then it applied IR technique to find document related to the topic. This approach uses disparate technologies, such as, IE for named entity extraction, association rule data mining, clustering of association rules, IR techniques, and few specific development that have wider applications.

The observation from experimental result leads to the conclusion that:

- Evaluating TopCat is difficult.
- TopCat is relatively insensitive to errors in named entity tagging.
- The segmentation of stories is more critical - since the result produced by TopCat is unreliable, if many documents contain multiple unrelated stories.

It raises two issues.

1) Is it possible to incrementally update new topics without looking at the old data?
2) How to make alert the user when new knowledge is being added to the topic?

An alternative approach to discover knowledge from text is to combine the information retrieval (IR) scheme (TF-IDF) for keyword/feature selection with association rule data mining technique for discovering knowledge (rules). Hany Mahgoub and et. al presented new text mining technique called, Extracting Association rules from text (EART) [2], for automatically extracting association rules from collection of text documents, the architecture of EART system is shown in figure 2.

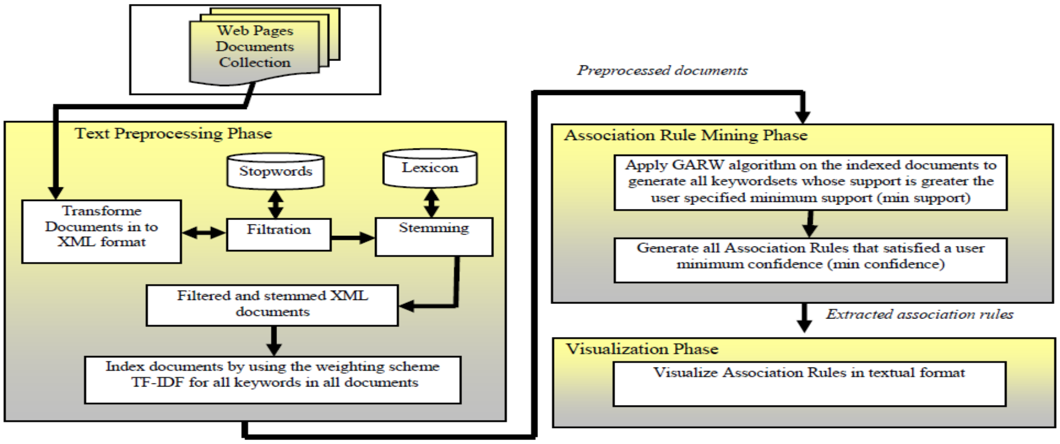
Web Pages
Documents
Collection
Preprocessed documents
Text Preprocessing Phase
Stopwords
Lexicon
Transforme Documents in to XML format
Filtration
Stemming
Filtered and stemmed XML documents
Index documents by using the weighting scheme TF-IDF for all keywords in all documents
Association Rule Mining Phase
Apply GARW algorithm on the indexed documents to generate all keywordsets whose support is greater the user specified minimum support (min support)
Generate all Association Rules that satisfied a user minimum confidence (min confidence)
Extracted association rules
Visualization Phase
Visualize Association Rules in textual format

Figure 2. Text Mining System Architecture.

The EART system integrates the XML technology with IR scheme TF-IDF for selecting the most discriminative keywords/features. The frequency based weighting scheme TF-IDF[2, 5, 8, 18] is used to index the documents by assigning higher weights to distinguished term in a document. The system then sort the keyword based on their weight score and selects only the top N frequent keywords up to M% of the number of running words. It then applied the own designed algorithm for **Generating Association Rules based on Weighting scheme** (GARW) for discovering most useful association rules. The GARW algorithm is as given below:

1. Let *N* denote the number of top keywords that satisfy the threshold weight value.

2. Store the top *N* keywords in index XML file along with their frequencies in all documents, their weight values TFIDF and documents ID. Four XML tags for all keywords (<doc-id>, <keyword>, <keyword-frequency>, <TF-IDF>) index the file.

3. Scan the indexed XML file and find all keywords that satisfy the threshold minimum support. These keywords are called large frequent 1-keywordset 1 *L* .

4. In $k \geq 2$, the candidate keywords *Ck* of size *k* are generated from large frequent (*k*-1)-keywordsets, *Lk*−1 that is generated in the last step.

5. Scan the index file, and compute the frequency of candidate keywordsets *Ck* that generated in **step** 4.

6. Compare the frequencies of candidate keywordsets with minimum support.

7. Large frequent *k*-keyword sets *k L*, which satisfies the minimum support, is found from step 6.

8. For each frequent keywordset, find all the association rules that satisfy the threshold minimum confidence.

Fatudimu IT and et. al developed a system similar to EART that is applied to per-election information collected from the website of the Nigerian Guardian newspaper [8]. The extracted association rules contain important features and describe the informative news included in the documents collection related to the concluded 2007 presidential election. The useful information presented by system could help to sanitize policy as well as to protect the nascent democracy.

5) Future Research:

Knowledge Discovery in text (KDT) plays an increasingly significant role in extracting interesting and non-trivial information and knowledge from unstructured text. In section 4, we discussed a technique called DISCOTEX for discovering knowledge form extracted data. It helps further to improve extraction by using this discovered knowledge. The predictive relationship between different slot fillers discovered by KDD is useful to predict additional potential extraction. For example, suppose the system discover the rule "VoiceXML ∈ language" ⟶ "mobile ∈ area". If the IE system extracted "VoiceXML ∈ language" but failed to extract “mobile ∈ area”, this assumes that there was an extraction error and add mobile to the area slot. This potentially improves the recall.

Most IE system is devolved using supervised approach. They either uses human annotated corpora or integrative framework to train the system. However constructing sufficient corpus for training accurate IE system is tedious and time consuming task. Further, interactive framework based system requires to provide good quality set of seed words to the system. One approach is to use automatic learning methods to decrease the amount of training data that uses vast repositories of annotated text. Siddharth patwardhan and et. al., presented an approach to exploits an existing set of IE patterns that were learned from small set of annotated training data to automatically identify new, domain-specific text from the web [9]. These web pages are then used for addition IE training, yielding a new set of domain specific IE patterns. However, more research is needed to explore methods for reducing the demand for supervised training data in IE.

Another approach to reduce the demand for supervised corpus-based training system is to develop unsupervised learning methods for building IE system. However, some work has been done in this area [5, 20]. This is another promising area for future research.

One of the most promising and emerging techniques in discovering knowledge from text is to combine the information retrieval schemes (such as TF-IDF) and data mining techniques (such as association rule mining) to discover knowledge (rules). We discussed some approaches [2, 5, 8] in brief in sec. 4. This approach can be further extended to use the concept feature to represent the text and to extract more usefulness that is more meaningful to represent the contents of the text document.

Another alternative approach to text data mining (TDM) is text knowledge mining (TKM). [14, 23]. The concept of tree-edit distance that allows not only the extraction of relevant text passages from the page of a given website, but also the identification of pages of interest and the extraction of a relevant text passages discovering the non-useful material is presented in [23].

6) Conclusion:

In this paper, we have reviewed several methods related to two approaches for mining text from text repositories using IE. In the first one, IE is used to extract abstract knowledge from text rather than concrete data. Regarding the first approach we discussed two methods in brief. One is related to keyphrase extraction, whereas the other focuses on building concept dictionaries for annotating a collection of documents from a particular domain.

In the second approach IE is used to obtain structured data from unstructured text and then KDD is applied to discover knowledge from this structured data. We mainly discussed two systems, viz. DISCOTEX and TOPCAT. Both use IE to view the problem in data mining/database context and then association rule mining technique is applied to discover the useful knowledge (rules). However, we also discussed other systems, EART and other similar one, that combines IR scheme TF-IDF (for keyword/ feature extraction) and association rule mining techniques for discovering knowledge. The key feature of EART system is that, the extracted association rules get the relations among the existing keywords in text documents collection ignoring the order in which these keywords occurs. It has been proved by the experimental observation that these techniques helps to improve the performance of the system by reducing execution time, on one hand, and extracts more interesting and useful rules than the rules generated by other systems. The experimental result of paper [6] also proved that the contribution of TF-IDF component in objective function enhance the result by improving precision from 58.86% to 71.45%, recall from 60.82% to 73.96% and F-measure from 59.82% to 72.68%.

➢ References:

[1] Stuart Rose and et. al. Automatic Keyword Extraction from Individual Document – Text Mining: Application and Theory – Michael W. Berry and Jacob Kogan- JOHN WILLEY & Sons, July 2010, ISBN-978-0-470-74982-1, pp 3-20.

[2] Hany Mahgoub, Dietmar Rosner, Nabil Ismail and Fawzy Torkey. A Text Mining Technique Using Association Rules Extraction- INTERNATIONAL JOURNAL OF COMPUTATIONALINTELLIGENCE VOLUME 4 NUMBER 1 2007 ISSN 1304-2386.

[3] Raymond J. Mooney and Un Yong Nahm. Text Mining with Information Extraction- *Multilingualism and Electronic Language Management:* Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005

[4] Mary Elaine Califf and Raymond J. Mooney (1997) Relational learning of Pattern Match Rules for Information Extraction. In T. M. Ellison (ed.) CoNLL97: Computational Natural Language Learning, ACL pp 9-15 1997

[5] Chris Clifton, Robert Cooley, and Jason Rennie TopCat: Data Mining for Topic Identification in a Text Corpus-IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, ISSN: 1041-4347/04 VOL. 16, NO. 8, AUGUST 2004

[6] Zhuoye Ding, Qi Zhang, Xuanjing Huang. Keyphrase Extraction from Online News Using Binary Integer Programming- Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 165–173, Chiang Mai, Thailand, November 8–13, 2011**.**

[7] Andrew Carlson, Justin Betteridge, Richard C. Wang. Coupled Semi-Supervised Learning for Information Extraction- WSDM’10, February 4–6, 2010, New York City, New York, USA. Copyright 2010 ACM ISBN: 978-1-60558-889-6

[8] Fatudimu I.T, Musa A.G, Ayo C.K, Sofoluwe A. B. Knowledge Discovery in Online Repositories: A Text Mining Approach- European Journal of Scientific Research ISSN 1450-216X Vol.22 No.2 (2008), pp.241-250

[9] Siddharth Patwardhan and Ellen Riloff. Learning Domain-Specific Information Extraction Patterns from the Web- IEBeyondDoc '06 Proceedings of the Workshop on Information Extraction Beyond The Document , Association for Computational Linguistics Stroudsburg, PA, USA 2006, Pages 66-73, ISBN:1-932432-74-4.

[10] Mirco Speretta and Susan Gauch. Using Text Mining to Enrich the Vocabulary of Domain Ontologies - 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 Pages 549-552 IEEE Computer Society Washington, DC, USA 2008, ISBN:978-0-7695-3496-1.

[11] Doug Downey, Oren Etzioni, Stephen Soderland, and Daniel S. Weld. Learning Text Patterns for Web Information Extraction and Assessment - Copyright © 2002, American Association for Artificial Intelligence, (www.aaai.org).

[12] Shantanu Godbole, Shantanu Godbole, Ajay Gupta. Building Re-usable Dictionary Repositories for Real-world Text Mining - *CIKM'10,* October 26–30, 2010, Toronto, Ontario, Canada. Copyright 2010 ACM 978-1-4503-0099-5.

[13] Seokhwan Kim, Minwoo Jeong, and Gary Geunbae Lee. A Local Tree Alignment-based Soft Pattern Matching Approach for Information Extraction - Proceedings of NAACL HLT 2009: Short Papers, pages 169–172, Boulder, Colorado, June 2009, Association for Computational Linguistic.

[14] D. S´anchez, M.J. Mart´ın-Bautista, I. Blanco, Text Knowledge Mining: An Alternative to Text Data Mining - 2008 IEEE International Conference on Data Mining Workshops, ISBN: 978-0-7695-3503-6.

[15] Tarik Zakzouk and Hassan Mathkour, Text Classifiers for Cricket Sports News - *2011 International Conference on Telecommunication Technology and Applications Proc .of CSIT vol.5 (2011) IACSIT Press, Singapore*

[16] Jamie Callan and Teruko Mitamura, KnowledgeBased Extraction of Named Entities - *CIKM'02,* Pages 532 – 537, November 4–9, 2002, McLean, Virginia, USA, ACM New York, NY, USA, ISBN:1-58113-492-4

[17] Huanzhong Duan, Yan Zheng, A Study on Features of the CRFs-based Chinese Named Entity Recognition- International Journal of Advanced Intelligence Volume 3, Number 2, pp.287-294, July, 2011.

[18] Manu Konchady, Text Mining application Programming- Cengage Learning Idia Private Ltd. ISBN-10-81-315-0247-3.

[19] Rogger Billisoly, Practical Text Mining With Perl- A JOHN WILLEY & Sons, INC Publication. ISBN 978-0-470-17643-6.

[20] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu Tal Shaked, Stephen Soderland, Daniel S.Weld, and Alexander Yates. Unsupervised Named-Entity Extraction from the Web: An Experimental Study – Artificial Intelligence Volume 165 Issue 1, Pages 91–134, June 2005, Elsevier Science Publishers Ltd. Essex, UK.

[21] Shady Shehata, Fakhri Karray, and Mohamed S. Kamel. An Efficient Concept-Based Mining Model for Enhancing Text Clustering - IEEE TRANSACTIONS ON

KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, Pages: 1360 – 1371, OCTOBER 2010,

[22] Shweta Mayor, Bhasker Pant. Document Classification Using Support Vector Machine - International Journal of Engineering Science and Technology (IJEST), ISSN : 0975-5462 Vol. 4 No.04 April 2012.

[23] Davi de Castro Reis, Paulo B. Golgher, Altigran S. da Silva, Alberto H. F. Laender - **Automatic Web News Extraction Using Tree Edit Distance**-WWW2004, May17-22,2004, NewYork,NewYork,USA. ACM1-58113-844-X