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, Issue 10, October 2014

A Study of Information Extraction Tools for Online English Newspapers (PDF): Comparative Analysis

M. Hanumanthappa[1], Deepa T. Nagalavi[2], Manish Kumar[3]

Professor, Dept. of Computer Science and Applications, Bangalore University, Bangalore, India[1]

Research scholar, Dept. of Computer Science and Applications, Bangalore University, Bangalore, India[2]

Dept. of Master of Computer Applications, M. S. Ramaiah Institute of Technology and Research Scholar, Dept. of Computers Science and Applications, Bangalore University, Bangalore, India[3]

ABSTRACT: Information retrieval is the task of retrieving relevant and useful information from e-newspapers. Electronic newspapers are electronic replicas of traditional newspapers. E-newspapers are becoming increasingly popular because of the ease and convenience in accessing them. Newspapers are the source of timely information. These are the documents comprising news items and several independent informative articles. It is also interesting to note that many newspapers present news on the same subject with different perspectives. In this fast moving era, it is impossible to read multiple newspapers. Thus, it is an essential to quickly summarize an article collected from different newspapers and present it to the reader in a compact and concise manner without compromising the structure and format of the news. A system that achieves this task should parse the e-newspapers available in PDF format and convert to text format. Secondly, data mining techniques are applied to identify and summarize the articles from various newspapers. This survey, focuses on article identification methods and popular extraction tools used for extracting the contents of e-newspapers for conversion from PDF to text format. A comparative study on extraction tools based on the source type, programming language and working characteristics is also presented.

KEYWORDS: PDF, Articles, E-Newspapers, graph clustering, Information Extraction.

I. INTRODUCTION

Information Retrieval (IR) is the activity of finding relevant information or a document that satisfies user information needs. The need for information retrieval arises on account of the volume of data. The traditional techniques of IR are not efficient because of the volume of data and hence there is a need for IR System. E-newspapers are widely read as compared to printed versions because they are not subject to geographic limitation. Throughout the internet, there are many e-newspapers available comprising many independent news articles. Retrieving relevant information from various e-newspapers is a difficult task. Information retrieval research focuses on retrieval of natural language from e-newspapers.

E-Newspapers are newspapers which are published electronically. They can take the form of normal print publications published on the Internet. In the present era e-newspapers play an important role providing information on current issues keep the readers up-to-date; also it plays a pivotal role in satisfying the information needs of many users. Newspaper pages are generally formed from several independent articles which are scattered throughout the page in columns. Identifying particular article is a relatively easy task for humans who can do it just by visual inspection, but however it is a difficult problem for a computer because e-newspapers have layouts which are not standardized and geometrically simple, and also previous approaches to the problem have not been robust. Data mining technique can be applied to efficiently identify and extract individual news items from e-newspaper; later the news items can be utilized in the data mining process to retrieve relevant information.

Data mining is the process of discovering interesting knowledge from large amounts of data stored in database, data warehouse, or other data repositories. Data mining has been popularly treated as synonym of knowledge discovery in database. In the literature various data mining techniques have been applied to mine articles from e-newspapers such as clustering and classification techniques. The clustering process consists of aggregating each block into sets of blocks (clusters). The clustering methods: K-mean, Agglomerative and genetic algorithms are used to cluster articles based on the textual content of the newspaper. Before identifying the textual content of an article, the newspaper has to transform into text format because the e-newspapers are basically PDF format documents. Hence PDF extraction tools can be used to convert PDF documents into text file.

Portable Document Format (PDF) is a platform independent file format which wraps many types of data such as images, texts, fonts, tables and many more. PDF documents can be viewed on any computer with the PDF viewer. The contents of document cannot be edited because the organization of data determined by the structure of PDF file. The file structure of PDF files consists of header, body, the cross reference table and trailer. Header includes a version number of PDF files, body is the main part of PDF file storing all types of objects, a cross reference table stores the location of each object stored in body part and trailer records the address of cross reference table. The document structure of PDF files can be represented as tree-like model which consists of many objects means that each node belongs to objects. The tree structure makes PDF document secure. Because of PDFs secure and wrapping features, the extraction of objects becomes a difficult task.

The Research Challenges of Information Retrieval System for identifying and retrieving articles from e-newspapers can be stated as under.

- Location: Identifying the location of relevant text from the document to extract the information.
- Acquisition and analysis: The layouts of news articles are not symmetrically identical hence extraction of relevant material, categorization and Clustering the columns is difficult task. Automated recognition and derivation of internal and crosswise relationships, especially to rule under construction.
- Vocabulary Mismatch: The same information is represented using a different vocabulary of words.
- Semantic analysis: If something really important happens, the news occupies more than one page of a newspaper and there are several related articles on each of these pages.
- Duplication: Articles taken from different e-newspapers are either near or exact duplicates. This estimation has not included the semantic duplicates yet.
- Heterogeneity of document: The contents present in e-newspapers are heterogeneous in nature, i.e. in addition to the text they might contain images, tables, and content based images.

The work focuses on the operations which run in two phases, one is to take text from e-newspaper of PDF format and the other one is to identify and extract articles from e-newspaper. The paper explores different data mining techniques, namely classification and clustering in order to automatically identify and extract the articles from e-newspapers. Further a survey is conducted on easily available information extraction tools to convert e-newspaper of PDF format to text format.

II. LITERATURE REVIEW

Information Retrieval System provides facilities to retrieve news items from e-newspaper. E-newspapers are the documents comprising several independent, informative articles which are scattered throughout the page in columns. The characteristics of a document object are its bounding box, its position within the page and its content. In the literature different techniques are applied to identify articles in newspaper.

Aiello and Pegoretti in [14]; worked on text processing techniques for the problem of article clustering in newspaper pages i.e. the identification of text blocks which belong to the same article. They proposed three text processing based algorithms namely simple clustering, comparative clustering and agglomerative clustering and these are called as graph clustering algorithms. An article objects is represented as the nodes of a graph, called the connection graph. An edge represents the fact that two objects belong to the same cluster. All the algorithms start with a graph with one node per document object and no edges. At each step one or more edges are added to the graph. At each iteration step of the algorithm, each connected sub-graph represents a portion of an article. The output of the algorithm is the graph in which each fully connected component represents a cluster, that is, a complete article. In all the three variants, a similarity threshold is fixed.

A similarity threshold is the degree of similarity between the two blocks and it is represented by the cosine of the angle between the weight vectors of two blocks. In the simple clustering algorithm (SCA) an edge in the connection graph is set by looking at the similarity matrix. For each element, if the value is above the threshold, then there is an edge in the graph. Whereas in the comparative clustering algorithm (CCA) the process of adding an edge is iterative and considers the edges present at each step in the graph. The CCA algorithm searches for the blocks which are more similar and then it compares the partially formed clusters of blocks before adding a new edge. The Agglomerative Clustering Algorithm (ACA) is to extend the motivation behind the Comparative Clustering. When two very similar blocks are found, the two blocks are linked not only via an edge, but are merged into a single block. This implies that after the merge, all the weights of the blocks need to be recomputed.

The clustering algorithms are evaluated by building a weighted harmonic mean of three distinct functions (precision, recall and distribution) that are computed comparing the ground- truth graph with the output graph. The ground truth for each page is the graph in which each article is a clique. Later authors compare the performance of the three algorithms determining that, Simple Clustering Algorithm has two main advantages: simplicity and efficiency.

Whereas the Comparative Clustering adds only a little complexity compared to the Simple Clustering algorithm; the exclusion of some edges in the connection graph should entail an improvement in the correctness of the edges set, with a possible loss in the portion of edges found. Finally, the Agglomerative Clustering Algorithm may improve both correctness and completeness, with a significant handicap: it is slower than the other two, since it must re-compute several times all the weights and the similarity matrix. However the simple clustering algorithm after removing stop words has high performance rate and a low computational complexity.

Based on the ground-truth graph approach [13], R. Beretta, L. Laura [1] proposed a methodology to evaluate newspaper article identification algorithms with the help of graph clustering techniques. The newspaper article identification problem is reduced to a specific graph clustering problem, i.e. a newspaper page is transformed into a graph, where each block is a node and all the nodes that belong to the same article are connected together. The algorithms are therefore evaluated using the appropriate coverage and performance measures that focus on intra-cluster density and extra-cluster sparsity i.e. the evaluation favors the algorithms that form a cluster in which most of the edges are from cluster to cluster.

The coverage of a graph clustering is the fraction of intra-cluster edges within the complete set of edges, and the performance counts the number of correctly interpreted pairs of nodes in a graph. The proposed methodology allows to easily distinguishing between different errors: the measures reward the correct detection of the bigger article with several blocks. Furthermore reversing the reduction means that turning a graph clustering algorithm into a full working newspaper article identification algorithm. An approach allows distinguishing different degrees of errors, depending on the number of blocks of the articles incorrectly identified.

Liangcai Gao, Zhi Tang, Xiaoyan Lin, Yongtao Wang [13] recovered the reading order of an article using the bipartite model. The bipartite graph model consists of two vertices one is predecessor and the other one is successor which is in line with bipartite graph matching. The vertices are the blocks within a page, the reading order of all blocks as an edge of the graph and the reading transition probability between the weights of the edges.

The common constraint of the document is that always the reading order will be top to down and left to right directions. Before selecting the linguistically portable blocks they first find the spatially admissible reading orders by spatial topology analysis. The reading transition score is calculated between text blocks by fusing several sources, including textual content, part-of-speech, position, style, and so on, and then optimal matching of the graph is matching with the maximum weight by taking the classic Kuhn-Munkres algorithm. With the reading order of all blocks on a page, article aggregation is segmenting the reading sequence into subsequences and merging them into articles.

Many different newspaper articles repeatedly contain duplicate information concerning the same event, but differ in choice of language used and sometimes with different perspective. Martina Naughton, Nicholas Kushmerick, and Joe Carthy in the work [16] focuses on merging descriptions of events from multiple sources to provide a concise description that combines the information from each source. The problem is decomposed into three sub-problems: (1) Annotation: identifying the spans of text in an article corresponding to the various events that it mentions; (2) Matching: identifying event descriptions from different articles that refer to the same event; and (3) Aggregation: converting the event descriptions into a structured form so that they can be merged into a coherent summary. The use of clustering technique is to automatically group sentences in terms of the events.

Sentence clusters are generated using average link, complete link and single link agglomerative clustering. Hierarchical agglomerative clustering (HAC) process start with each data point to a singleton cluster, and then repeatedly merges clusters until there is no more non-clustered element left. HAC clustering methods require a similarity metric between two sentences. The standard cosine metric is used over a bag-of-words encoding of each sentence where all stop words are removed. Authors have developed Term Frequency and Inverse Document Frequency (TFIDF) like weighting scheme where document is defined to be the set of sentences which discuss a given event and then weight terms according to their frequency in the document compared to the entire corpus. But an approach did not use the term weighting.

Prof. A.D. Thakare, N. Muthiyan, D. Nangade, D. Patil, M. Patil [17] proposed a system that uses the capability of Genetic Algorithm (GA) to form the clusters of news articles. The approach is to find potential and hidden knowledge from the repositories for decision making document clustering of news articles using Genetic Algorithm. Newspaper articles will be grouped into different domains such as business, economics, politics, sports, entertainment, social, environmental, etc. according to the similarity of the keywords using an algorithm. The system contains set of newspaper articles, and the articles are clustered based on similarity of keywords into different domains. Genetic Algorithms perform the same operations on the population of possible targets with only those that fit the solution better surviving.

Genetic Algorithm consists of four elements: The first is the population of chromosomes, which represent the possible solutions of the problem. Selection is the second element and it refers to the part of the population that will evolve to the next generation. Selection is performed based on a fitness function. The selection process is applied to each generation produced. Crossover refers to the combination or exchange of characteristics between two members of the group defined by selection, by which offspring is produced. The process repeats until optimized clusters are formed. Genetic Algorithm is compared with K-means clustering algorithm. K-means is used for local optimization where as Genetic Algorithm is used for global optimization. Genetic Algorithm can generate better results than k-means. More optimized clusters are formed using Genetic Algorithm. It is Hybrid Model and it can do automatic clustering. It is used for Better searching techniques

III. INFORMATION EXTRACTION TOOLS

E-Newspapers are electronic documents available in the form of PDF files. In literature many tools are introduced which are employed to pull data from PDF files. There are many other easily available extraction tools used to convert to text format.

- PDFBox: Apache PDFBox [4] is a library which consists of the parser classes, extraction and basic modification classes. A parser of PDF file first locates the trailer and then obtains the address of cross reference table, later it accesses each node from a tree like model to get information. These functions are implemented by a PDF class library called PDFBox. Fang Yuan et al [3] used it for extracting title, author, address, abstract, keywords and the class number of the document. In this work authors extract text using PDFBox and inject tags wherever format change happens to convert the document to semi structure format. This project allows creation of new PDF documents, manipulation of existing documents and the ability to extract the contents of documents.

- PDFtohtml: PDFtohtml [5] is a utility which converts PDF files into HTML and XML formats. Burcu Yildiz, Katharina Kaiser, Silvia Miksch [2] used it to convert PDF to XML to extract tables from a PDF document, here the

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, Issue 10, October 2014

tool returns text chunks and their absolute coordinates in the PDF file in the same order as they are inserted into the original file. PDFtohtml is a tool based on the open source viewer XPDF. The commercial application is available only in executable format. Xpdf [7] is an open source viewer for Portable Document Format (PDF) files. The Xpdf project also includes a PDF text extractor, PDF-to-PostScript converter, and various other utilities. Xpdf should work on pretty much any system which runs X11 and has Unix-like (POSIX) libraries. It requires ANSI C++ and C compilers to compile it. The main problem with this tool is that it is tough to extract images.

- PDFLlb TET: PDFLib TET [9] (Text Extraction Toolkit) is a tool used to extract text, images and metadata from PDF documents. TET extracts text contents of a PDF as Unicode strings, along with the detailed glyph and font information as well as the position on the page. It also converts PDF documents to an XML-based format called TETML which contains text and metadata as well as resource information. TET contains advanced content analysis algorithms for finding word boundaries, grouping text into columns and removing redundant text. PDFLib TET can be used to implement the PDF indexer for a search engine, to repurpose the text and images in PDFs, to convert the contents of PDFs to other formats and to process PDFs based on their contents.

- Solid-PDF Tools: Solid Convertor [11] is a document reconstruction software product which allows the users to convert PDFs into editable documents. The software creates PDFs from a variety of file sources by saving the original format. The problem with the tool is that it is only limited to the windows platforms and image extraction is a hard job.

- iText: Developers will use iText [8] to serve PDF to a browser, generate dynamic documents from XML file or databases. The tools make use of PDF’s many interactive features and it splits, concatenate and manipulate PDF pages. iText is also used to automate filling out PDF forms also to add digital signatures to a PDF file. In [18] authors done survey on different PDF extraction tools and came up with a result that compare to other extraction tools iText tool is the best tool to extract all objects and its information.

- 3-Heights PDF Extract: PDF Extract tool [10] is used for reading out the contents and properties of PDF documents. Also, it extracts the contents quickly and efficiently. 3-Heights tool extract text as Unicode from the character, word or page also it supports text which does not contain blank characters. It searches for keywords and retrieve their position. This tool can be utilized to convert PDF document to text document.

- Able2extract: Convert PDF files to popular MS Office formats including Excel, Word, PowerPoint and Publisher, as well as non-Microsoft formats such as AutoCAD. Able2Extract 8 is a fully cross-platform solution, available for Windows, Mac and Linux users. Editing, analyzing, and re-versioning PDF documents.

- Aspose: Aspose converts PDF to word processing documents. Aspose reads a PDF, recognizes editable document structures and allows saving the resulting document as DOC, OOXML, ODT, RTF, WordML, HTML, MHTML or TXT. The Aspose tool converts the contents exactly including format of the original document.

IV. COMPARISON

- The objective of this comparative study is to search for a tool which converts e-newspapers of PDF format into text format retaining the original layout of the document. Later text mining techniques are applied to analyze news items. Hence comparisons between the different tools are identified based on their universal characteristics and operating features. The tools listed in table1 are the tools users use to convert PDF files to other formats. Whereas the tools listed in table2 are the library files used by developers to add and create PDF features.

Table 1: Tools to convert PDF files to other formats

Name	License	Platform	Language	Extracts				Other Description
				Image	Text	HTML	Font	
PDFBox	Apache License	Linux, Unix, Windows	Java, C#	√	√	√	--	Creating pdf, bookmarking pdf
Solid Converter PDF	Proprietary	Windows	--	--	√	√	√	Converts PDF to Word, Excel, supports passwords and batch conversion
Able2extract	Proprietary	Windows, MacOSX, Linux	--	√	√	√	--	Converts PDF to Word, Excel, PowerPoint, Publisher, Open Office, AutoCAD
PDFtohtml	Open Source	Linux, Unix, windows	C++	--	√	√	√	Converts PDF to XML and png images
Aspose	corporate	Linux, Unix, windows	Java, .NET	√	√	√	√	Converts to DOC, OOXML, ODT, WordML, RTF, MHTML

The tools listed in the table1 and table2 are used to extract the data from pdf files and convert it into text format. The iText, Jpedal, Aspose and PDFLib tools are having more features as compared to other tools. Among these tools Aspose tool is the best extraction tool as it converts the document along with the layout, whereas iText, Jpedal, PDFLib tools converts the documents without structure and disadvantage with these tools is that the text's reading order is not necessary preserved, especially when handling multi-column documents with complex layout.

Table 2: Libraries to add and create PDF features

Name	License	Platform	Language	Extracts				Other Description
				Image	Text	HTML	Font	
iText	AGPL	Linux, Unix, Windows	Java, c#, .NET	√	√	√	√	Open-source library to create and manipulate PDF, RTF
JPedal	Proprietary, GNU LGPL	Linux, windows	java	√	√	√	√	A Java developer library to view, extract, print PDF files
Aspose	corporate	Linux, Unix, windows	Java, .NET	√	√	√	√	Converts to DOC, OOXML, ODT, WordML, RTF, ,MHTML
PDFLib TET	Commercial	Linux, Unix, windows	Multi language	√	√	√	√	metadata from PDF documents.
3-Heights PDF Extract	Commercial	Linux, Unix, windows	Multi language	√	√	--	√	Process data in forms

- **THE COMPARITIVE STUDY OF LITERATURE WORK TO IDENTIFY AND EXTRACT ARTICLES:**

The authors Aiello &Pegoretti [14] have evaluated three algorithms for article identification. They build a graph called Connection graph in which each node is a block. Each connected component is an article. The ground truth for each page is the graph in which each article is a Clique. The function Weighted Harmonic mean is computed by comparing the ground-truth graph with the output graph. The ground truth base is made of one clique for each article. The algorithms identify the articles as the connected components of the connection graph and evaluate the algorithm with respect to how closely the connection graph matches the ground truth graph. The method demonstrated the benefits of semantic information, but it fails when several independent articles share same textual content. One more disadvantage is that, the Approach is used only to evaluate the algorithms that build a graph structure.

Whereas the authors Beretta & Laura have mentioned in [1] that the nodes of the graph are the blocks and implemented an evaluation metric which includes the operations like loading a set of PDF files with their associated ground truth base, Defining and saving a ground-truth base for a PDF file and Evaluating the performance of algorithms. The newspaper article identification problem is reduced to a specific graph clustering problem, after evaluating the algorithm the reduction is reversed into a full working newspaper article identification algorithm. An approach allows distinguishing different degrees of errors, depending on the number of blocks of the articles incorrectly identified.

In [1] and [14] authors used clustering technique to group the text blocks belonging to the same article and this demonstrates the semantic information for the newspaper document understanding. Whereas the authors Liangcai Gao, Zhi Tang, Xiaoyan Lin, Yongtao Wang in [13] recovered the method for article reconstruction. The bipartite graph model is used to detect the reading order of an article based on content similarity and spatially admissible reading orders by spatial topology analysis. The corresponding optimal matching is obtained which generates one or more block queues. An article is segmented into sub-queues then generated sub-queues are merge into articles.

Prof. A.D. Thakare, N. Muthiyan, D. Nangade, D. Patil, M. Patil [17] used Genetic algorithm to form the clusters of news articles. The articles are grouped based on its features where it first searches the keywords then the articles are clustered based on similarity of keywords into different domains. The disadvantage with this algorithm is that the documents are clustered based on similarity of keywords only because this approach fails when several independent articles share same textual content.

V. PROPOSED PLAN OF WORK

The E-newspapers from different news sites are basically available in the PDF format. Hence the PDF extraction tool is applied to extract text from e-newspaper without affecting the format of a newspaper. Later text mining techniques are applied to analyse news items and summarize the information to readers. Thus article identification methods are applied to extract articles from newspapers. The important characteristic of an article is its bounding box, its position within the page, and it's content. The main concentration is on the contents of an article. In newspaper page all the blocks of text in which the layout divides the articles together with their textual content. A method to state whether two blocks belong to the same article is to compare the words they contain if they share the same words they are likely to be about the same subject. Then similarity scores based on frequency of individual terms are calculated using frequency data such as vector space method and probabilistic model. The article clustering process consists of four steps: i) Obtain the list of all the words inside the blocks. ii) Give a weight to each word inside each block. iii) Find the similarity between all the pairs of vectors. iv) Group together the blocks which probably belong to the same article. v) Find the reading order of blocks and merge the blocks into article of sequence. Furthermore the identified articles are extracted from each of the newspapers. The extracted articles from different newspapers will be grouped into different domains such as business, economics, politics, education, sports, etc. according to the similarity of the keywords. Furthermore the text mining techniques are applied to analyze the news articles.

VI. CONCLUSION

E-newspapers play a substantial role in providing the useful information to users. Each newspaper contains several independent, informative news articles scattered on the page and have layouts which are not standardized and not geometrically simple. This paper has explored the data mining techniques that the researchers have proposed and experimented earlier for identifying articles from e-newspapers. It is evident that in order to analyze news, article identification is important. It is also observed that clustering algorithms such as graph clustering algorithm or genetic algorithm can be applied to identify articles. Once the article is identified, it can be used to summarize the information. The newspapers, collected from news sites are basically in PDF format and text mining from these formats is not an easy task. Various extraction tools which are listed in the paper are studied in order to find best suitable extraction tools to convert the data from PDF to text format.

REFERENCES

[1] Beretta, R., Laura, L., "Performance Evaluation of Algorithms for Newspaper Article Identification" Published in: Document Analysis and Recognition (ICDAR), IEEE International Conference on DOI : 10.1109/ICDAR.2011.87 , 2011

[2] Burcu Yildiz , Katharina Kaiser , Silvia Miksch "pdf2table: A Method to Extract Table Information from PDF Files" http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.94.9382 , 1773-1785, IICAI 2005

[3] Fang Yuan and Bo Lu, "A new method of information extraction from PDF files" , Published in: Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on (Volume:3) DOI : 10.1109/ICMLC.2005.1527225

[4] http://pdfbox.apache.org/

[5] http://pdftohtml.sourceforge.net/

[6] http://www.adobe.com/enterprise/standards/

[7] http://www.foolabs.com/xpdf/

[8] http://www.itextpdf.com/book/

[9] http://www.pdflib.com/products/tet/

[10] http://www.pdf-tools.com/

[11] http://www.soliddocuments.com/

[12] K. H. Talukder, Md. Mustaneer Rahman, T. Ahmed, "An Efficient Speech Generation Method based character and modifier of Bangla PDF Document", IEEE, Proceedings of 13th International Conference on Computer and Information Technology, ICCIT 2010.

[13] Liangcai Gao, Zhi Tang, Xiaoyan Lin, Yongtao Wang, "A Graph-based Method of Newspaper Article Reconstruction", 21st International Conference on Pattern Recognition (ICPR 2012), Tsukuba, Japan, November 11-15, 2012.

[14] M. Aiello , A. Pegoretti , "Textual Article Clustering in Newspaper Pages " , Applied Artificial Intelligence, vol 20, no 9, pp. 767-796, 2006, [online] available: http:// dx.doi.org/10.1080/08839510600903858

[15] M. Minami, H. Morikawa, T. Aoyama, " The Design of Naming-based Composition System for Ubiquitous Computing Applications", 2004 IEEE, Proceedings of the 2004 International Symposium on Applications and the Internet Workshops.

[16] Martina Naughton, Nicholas Kushmerick, and Joe Carthy," Clustering sentences for discovering events in news articles", ECIR 2006: 535-538

[17] Prof. A.D. Thakare,N. Muthiyan, D. Nangde, D Patil, M. Patil, "Clustering Of News Articles to Extract Hidden Knowledge" IJETAE Webs e: www.ijetae.com , SSN 2250-2459, Volume 2, Issue 11, November 2012.

[18] Sarang Pitale,Tripti Sharma, "Information Extraction Tools for Portable Document Format " Sarang Pitale et al, Int. J. Comp. Tech. Appl., Vol 2 (6), 2047-2051, 2047 ISSN:2229-6093

BIOGRAPHY

Dr. M Hanumanthappa is currently working as Professor in the Department of Computer Science and Applications, Bangalore University, Bangalore, India. He has over 18 Years of teaching (Post Graduate) as well as Industry experience. His area of Interest includes mainly Data Mining, Information Retrieval and Programming Languages. Besides, He has conducted a number of training programmes and workshops for Computer Science students. He is also the Principle Investigator of UGC-Major Research Project; he has published nearly 50 Research Papers in National and International Journal and Conferences. Currently he is guiding students for PhD in Computer Science, under Bangalore University. He is also one of the members of Board of Studies as well as Board of Examiners for various Universities of Karnataka.

Deepa T. Nagalavi received MCA degree from Karnataka University, Dharwad, India in 2008. Currently she is perceiving her PhD in Computer Science (Data Mining) from Bangalore University Bangalore, India. She has over 3 years of teaching experience. The areas of interest are Information Retrieval, Data Mining and Programming Languages.

Manish Kumar is working as Assistant Professor in Department of Master of Computer Applications, M. S. Ramaiah Institute of Technology, Bangalore, India. His specialization is in Network and Information Security and Computer Forensic. He has worked on the R&D projects related on theoretical and practical issues about a conceptual framework for E-Mail, Web site, and Cell Phone tracking, which could assist in curbing misuse of Information Technology and Cyber Crime.