International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 1 (February '2017)
Special Issue on 1st National Conference on Advanced Computing and Communication Technologies,
Conference held at Siri PSG Arts and Science College for Women, Sankari,Tamilnadu,India.

THE EXTRACTION OF ONLINE NEWSPAPERS FOR CLASSIFICATION OF INDIVIDUAL NEWS ITEMS

K.Veena,
Assistant Professor,
Department of Computer Science,
J.K.K.Nataraja College of Arts and Science,
Kumarapalayam, Tamilnadu, India.

R.Malathi,
Assistant Professor,
Department of Computer Science,
J.K.K.Nataraja College of Arts and Science,
Kumarapalayam, Tamilnadu, India.

Dr.V.Sasirekha,
Assistant Professor,
Department of Computer Science,
J.K.K.Nataraja College of Arts and Science,
Kumarapalayam, Tamilnadu, India.

Abstract: Newspapers are widely used information source for the past century. The electronic media converts the newspaper as web pages. The web newspapers are updated frequently with respect to the news collection and its importance. The text mining techniques are used to analyze the web newspapers. The page mining techniques are currently used to extract the web newspapers. This paper proposes the process of automated page extraction and classification by using complete page mining mechanism. This mechanism is divided as four phases: download, extraction, cleaning and classification. The download phase downloads the entire newspaper pages into the local machine. The extraction process extracts the news item from the stored pages. The irrelevant and redundant data are removed in the cleaning phase. The classification phase arranges the news item documents based on the content using knowledge base based classification technique. The extraction and classification can be applied the news item from any web newspaper. Reduction in storage requirements, automation of news item extraction and fast information retrieval are the major features of the system.

Keywords: Newspaper, Extraction, Cleaning, Classification.

I.INTRODUCTION

Newspapers are usually printed on inexpensive, off-white paper known as newsprint. Since the 1980s, the newspaper industry has largely moved away from lower-quality letterpress printing to higher-quality, four-color process, offset printing [1]. In addition, desktop computers, word processing software, graphics software, digital cameras and digital prepress and typesetting technologies have enabled newspapers to publish color photographs and graphics, as well as innovative layouts and better design. Newspapers have been developed around very wide topic areas, such as news for merchants in a specific industry, fans of particular sports, fans of the arts or of specific artists, and participants in the same sorts of activities or lifestyles. It includes political events, crime, business, culture, sports, and opinions, photographs to illustrate stories, cartoons to illustrate the opinion, weather forecasts, an advice column, critic reviews of movies, plays and restaurants, etc, editorial opinions, a gossip column, comic strips, entertainment contents like crosswords, sudoku and horoscopes, a sports column or section, a humor column or section and a food column [2]. One advantage to newsprint is that reading it does not require any sophisticated, cumbersome technical equipment. This offers the reader a high level of flexibility: can be read in any place at any time.

On-Line Newspaper:

As a form of computer-aided communication, the World Wide Web (WWW) is equally ambivalent for the print media. Its technical potential greatly surpasses that of the printed newspapers in a number of ways [3]. WWW has the advantages of being interactive, multimedia, of providing internal and external networks and offering selection functions, the possibility of regular updates, access to archives, rapid access to a large number of newspapers, and being paperless, thus creating no problems of waste disposal. On the one hand, WWW presents a threat to the traditional distribution system [4]; on the other hand it gives publishing houses the opportunity to offer up-to-date information, advertisements and additional services via a further communication channel. The disadvantages of online newspapers can not provide the same experience of reading as a print newspaper, the download times are long, and the "paper" lacks portability. In case of technical problems connected with reading online newspapers and the cost of Internet access.

Data Mining Concepts:

Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries

[5]. Several types of analytical software are available: statistical, machine learning, and neural networks.

- **Classes**: The stored data is used to locate data in predetermined groups.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences.
- **Associations**: Data can be mined to identify associations.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system [6].
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Classification:

The classification approach uses a pre-existing ontology as reference ontology. Each Web page from the local site to be characterized is automatically classified into the best matching concept in the reference ontology. The documents that have been manually attached to each concept in the reference ontology are used as training data for categorizing the new documents. A vector space approach is used for calculating the cosine similarity measure to identify the closest match between a vector representing the Web page and the vectors representing the concepts in the reference ontology [7]. All the training documents associated with a given concept are concatenated to form a super document. These super documents are then indexed using a vector space retrieval engine. Each Web page to be categorized is then treated as a query and the retrieval engine returns the top matching super documents for that "query". Each document is attached to only the top matching and the weights of all the document matches to a particular concept are then accumulated to determine the weight of that concept in the site being studied.

IEPAD: Web Information Extraction based on Pattern Discovery:

In recent years, collation of information on the World Wide Web has become the main issue in applications that provide value-added services. Extracting data from the Web requires a wrapper that "wraps" around a Web site and extracts data from Web pages. Wrapper induction is one of such techniques that learn extraction rules by generalizing from training examples. The key idea underlying these wrappers is that the information to be extracted can be located based on "landmarks". That is, the wrappers extract a tuple by scanning the input Web page, recognizing the "landmarks" that are used to delimit information records. Kushmerick etal, identified a family of wrapper classes and the corresponding induction algorithms which generalize from labeled examples to extraction rules. More expressive wrapper structures are introduced lately [8]. Softmealy by Hsu and Dung uses a wrapper induction algorithm to generate extractors that are expressed as finite-state transducers. Meanwhile, Muslea et al. proposed "STALKER" that generates wrappers based on a set of disjunctive landmark automata organized as a hierarchy. The PAT tree patterns in a given input string can be efficiently identified. A PAT tree is an efficient data structure successfully used in the area of information retrieval for indexing a continuous data stream using this data structure to index an input string, all possible character strings, including their frequency counts and their positions in the original input string can be easily retrieved.

II.IMPLEMENTATION

The system is designed as graphical user interface applications which are carried out in two modes. The newspaper download process is done under the online mode. All the cleaning and classification activities are performed over the offline mode. The analysis tasks are done with the downloaded documents that are stored in the local machine. The system is divided into four major modules. They are download module, extraction process, cleaning process and classification process. The download module is designed to download the newspaper contents from the web sites. The extraction process is designed to collect the news items from the downloaded pages. The document-cleaning module is designed to clean the newspaper documents that are downloaded in the local machine. The classification module is designed to classify the newspaper documents with reference to the knowledge base.

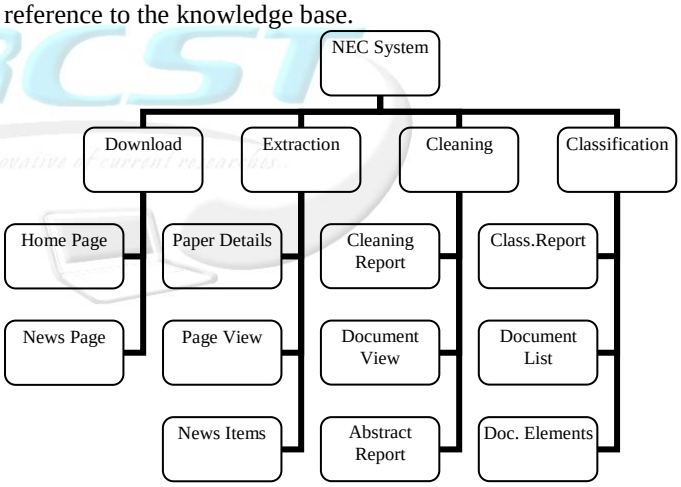

Figure 1.1 System Architecture

Download Process:

The download process is the initial interface for the system in which newspapers that are available on the Internet is showed in the list of newspapers and its URL details. The user can choose any newspaper from the list. The downloading process gets started after the newspaper selection. The user can add additional newspapers into the database directly. The newspaper downloading process is done with the support of the Transmission Control Protocol (TCP) and Hyper Text Transfer Protocol (HTTP). The newspaper download process is performed with two phases. They are home page download process and the news pages download process. The home page download process is applied to download the home page for the newspaper. Each newspaper maintains a home page to integrate all news

items with various categories. All the related news items are presented in separate pages. The news item pages are connected with the main page by using the hyperlinks. All the page navigation activities are carried out with the help of the hyperlinks. The home page is applied into parsing process to fetch links from the home page. This task is achieved by using the Hyper Text Markup Language (HTML) tags. The anchor tag is used to extract links from the home page. The link list is transferred into the relevant link identification process. In this process the links that refers external page are eliminated from the link list. The final list contains the relevant links only. The final link list denotes the news item pages. The news pages are downloaded from the web site. The news pages are stored in to the home directory for the news page. The downloaded process considers only the text elements. The images and other multimedia contents are not considered in the download process. The download process details are updated into the database. The news pages are identified with a unique document identification number. All the cleaning and classification process are done with the downloaded new pages.

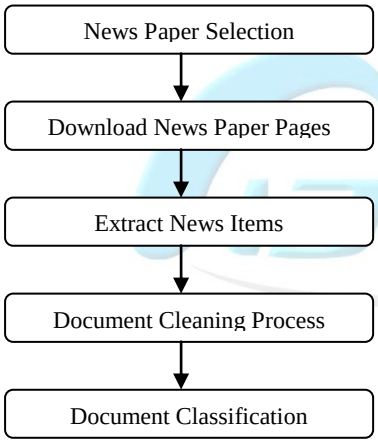

Figure 1.2 System Flow Diagrams

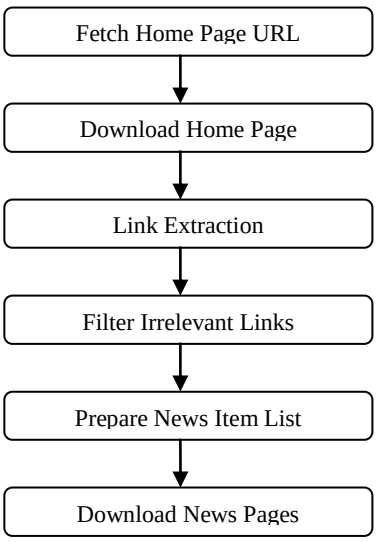

Figure 1.3 Newspaper Download Process

Extraction Process :

The extraction process is designed to extract the news items from the news pages. This process is applied on the home page contents. The list of words those are associated with the hyperlinks for the news pages are considered as news items [9]. The news items and the links are stored into the database. All the extraction activities are initiated from the home page window. The extraction process results are updated and retrieved from the database.

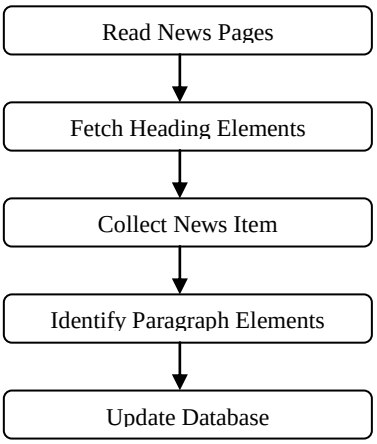

Figure 1.4 News Item Extraction Process

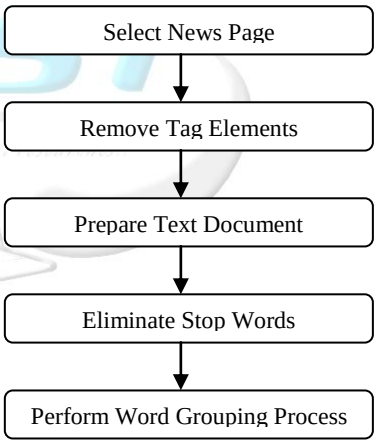

Figure 1.5 Document Cleaning Process

The extraction process is divided into three sub modules. They are paper details, page view and news items list. The paper details show the list of news pages with its size information. The user can view the page content in a manipulated format. The news items are extracted from the home page is listed in the news items list. All the three modules can be run under the offline environment. The paragraph elements and also the heading elements are also separated from the document.

Cleaning Process:

The document cleaning process is designed to clean the news pages by removing the tags that are used in the pages. The cleaning process is done in two ways: tag elimination and word analysis [11]. The tag elimination process is

performed on the downloaded news pages. The content of the news page is extracted from the web page without the tag elements. The output of the tag elimination process is stored into a text document. The word analysis process is carried with two major operations. They are stop word elimination and the word counting process. The system maintains a stop word collection and all the documents are verified with the stop words collections. The stop words that are located in the text documents are removed from the document. The remaining words are updated into the database with the occurrence. The system provides three types of reports in the cleaning process. They are the document cleaning report, document view and the document abstract report. The document cleaning report shows the list of documents and their total words and relevant words count. The user can view the cleaned document in the document view. The document abstract report presents the list of words and their count for the selected document.

Classification Process:

The document classification process is designed to classify the documents based on the document content. The document classification refers the process of grouping the documents. Automated document classification a method does not requires any knowledge base for the classification process. They perform the learning process to update the knowledge base. This system uses a knowledge base for the classification process [10]. The knowledge base maintains a collection of words with the group name and the sub group name. The knowledge base is constructed manually by the user. The classification strength is measured by the knowledge base. The user can update the knowledge base with the group and sub group names. The system uses the cleaned document for the classification process. The words that are extracted from the news page documents are compared with the stored knowledge base. The matched word and its group name are identified. The count for the matched words and the group are estimated and the maximum count is retrieved from the comparison process. The document is classified with the relevant group name. The system uses the cleaned document for the classification process. The words that are extracted from the news page documents are compared with the stored knowledge base. The matched word and its group name are identified. The count for the matched words and the group are estimated and the maximum count is retrieved from the comparison process. The document is classified with the relevant group name. The classification process has three sub modules. They are classification report, document list and document elements. The classification report shows the group name, sub group names with the document count. The document list provides the list of documents in each group and sub group. The document elements report presents the paragraph element and heading elements for each document.

III. SCOPE FOR FUTURE DEVELOPMENT

The future development of the system can include the following features.

- The classification process in the current system is done with the support of the predefined knowledge base. The knowledge base maintains a set of groups and sub group collection. The future development of the system can automate the classification based on the dynamic learning mechanism.
- The data mining techniques like clustering and association rule mining techniques can be added in the future development of the system.
- The system can be enhanced with the text summarization techniques to fetch the summary of the newspaper and its news items.
- The redundant data elimination mechanisms can be used remove the redundant news items in the newspapers.
- In the current system the noise data elimination are performed with the line elements in the newspaper documents. The future development of the system can include the web structure mining techniques to remove the irrelevant data items.
- The neural networks, fuzzy logic and Bayesian classification techniques can be used for the document classification process.

IV. CONCLUSION

The "News Item Extraction and Classification system" is developed to mining and classify the news items in the online newspapers, downloads the entire newspaper websites to the local machine and techniques are applied into the local document repository. This system is tested with the various newspapers websites. The system produces the intermediate and final results in various forms. All the results are finally verified with the produced output. This system delivers the better performance for the downloading and classification tasks. Two additional aspects of news item extraction should also be mentioned: 1) the removal of non-relevant formation also makes news item extraction useful as a way of data cleaning, and 2) the space needed for storing the news items is much smaller than what is needed for the pages themselves, typically only 10% of the original size, and this means that the news items can be used as summary information for web sites and stored in a web warehouse. The classification simplifies the news item retrieval process in an effective way.

V.REFERENCES

[1]. Christoph Neuberger, Jan Tonnemacher, Matthias Biebl and André Duck, "Online--**The Future of Newspapers" Germany's Dailies on the World Wide Web",** Department of Journalism, Catholic University of Eichstatt- 1998.

[2]. Shelly Rodgers, Glen T. Cameron and Ann M. Brill, Ad Placement in E-Newspapers Affects Memory, Attitude, Newspaper Research Journal (2005)

[3]. Markus Zinnbauer, Ludwig-Maximilians, "**e-Newspaper: Consumer Demands on Attributes and Features",** University of Munich, Germany – 2003.

[4]. Helena Ahonen, Oskari Heinonen (1997) **"Applying Data Mining Techniques Text Analysis",** Report C-1997-23, Department of computer Science Univrsity of Helsinki 1997

[5]. Arun K.Pujari, **"Data mining Techniques"** University Press, First Edition, 2001.

[6]. **Prabhu C.S.R, "Data warehousing Concepts, Techniques, Products and Applications".** Published by Ashoke K.Ghosh, Prentice-Hall of India Private Limited, M-97, Connought Circus, New Delhi-110001. First Edition 2001
[7]. Shian-Hua Lin and Jan-Ming Ho, "**Discovering Informative Content Blocks from Web Documents",** Institute of Information Science, Academia Sinica, Taiwan.
[8]. Chia-Hui Chang and Chun-Nan Hsu, "**IEPAD:Web_Information Extraction based on Pattren Discovery"** In Proceedings 1999 National Computer Symposium, Tamkang University, Tamsui, Taiwan.1999.
[9]. Deng Cai1, Shipeng Yu, **"Extracting Content Structure for Web Pages based on Visual Representation"**, Microsoft Research Asia
[10]. Karel Fuka, Rudolf Hanka, **"Feature Set Reduction for Document Classification Problems",** Medical Informatics Unit, University of Cambridge.
[11]. Hung-Yu Kao, Shian-Hua Lin, **"Mining Web Informative Structures and Contents Based on Entropy Analysis",** National Taiwan University, Taipei, Taiwan, ROC.
[12]. Jiawei Han and Micheline Kamber, **"Data Mining: Concepts and Techniques",** Kaufmann Publishers, 2000.
[13]. Karel Fuka, Rudolf Hanka, **"Feature Set Reduction for Document Classification Problems",** Medical Informatics Unit, University of Cambridge.
[14]. Simon Henriksson, Mats Lindqvist and Martin Söderblom, **"E-newspaper Navigation",** School of Information Science, Computer and Electrical Engineering, Halmstad University.
[15]. http://www-i5informatik.rwthaachen.de/
[16]. http://people.howstuffworks.com/newspaper1.html
[17]. http://en.wikipedia.org/wiki/News_values
[18]. http://www.editorsweblog.org/2005/07/onlinenewpaper.php
[19]. http://jcmc.indiana.edu/vol4/issue1/neuberger.html
[20]. http://www.firstmonday.org/issues/issue8_9/nguyen/
[21]. http://www.ru.ac.za/library/infolit/news.html
[22]. http:// www.sims.berkeley.edu/ ~hearst/ text-mining.html,