AN EXPERT SYSTEM MODEL THAT ENABLES THE DEVELOPMENT OF INSTITUTIONAL KNOWLEDGE USING TEXT MINING METHODS IN OPEN SOURCE SOFTWARE

[1]AYTUG BOYACI, [2]MUSTAFA ULAS

[1,2]Firat University
E-mail: aytugboyaci@gmail.com, mustafaulas@gmail.com

Abstract- With the rapidly developing information technology, institutions store the most important data in digital storage. This information is the most important data sources to create the company's institutional knowledge. All information obtained by the Company is the company's know-how. There are many workflow process such as reducing the cost, improving the quality of products and services offered by the company, ensure the efficiency, satisfying the expectations and demands of customers. There are several open source applications to manage each of these workflow processes and these applications produce qualitative or quantitative data. Management of big data which increases with growth of the company is quite difficult. This may cause false recognition of the important knowledge which is hidden in big data while data analysis and reporting. Text mining methods can be used for the analysis which is extremely important in order to increase the enterprise business intelligence of these data. In this study, A model is suggested to generate useful knowledge in big data which created by open source applications with text mining methods such as classification, association analysis, feature extraction and clustering.

Keywords- Big Data, Text Mining, Expert Systems, Open Source Software, Institutional Knowledge.

I. INTRODUCTION

With the rapid development of information technologies, corporations store most of the data digitally integrating them to computer systems. It is highly important to transmit the professional experience which is gained in years for the development of corporations. This experience builds the institutional memory. To store the documents and data in archives contributes to build an effective institutional memory [1][2].

A great deal of open source business software are used to configure and analyze the documents and data of corporations. These software generally obtain access to documents or report due to the content or some features of it. Also these software help to record the institutional memory with some features such as access and update utilities, avoiding data repeat and useless data mass [2][3].

Despite the digitalization and analyzing of data speeds up the development of institutional memory which is built in years, it is an important problem that this development momentum depends on the reports which can be produced by used application software and workforce of human. It gets harder to analyze the increasing amount of data as the corporation gets bigger. For this reason, providing the faster development of institutional memory of corporations is based on a working model which is independent of human workforce and software. This can only be realized with the analyzing of all working procedures by an expert system. By designing an expert system model, it is possible to reach the data which can't be foreseen and can be analyzed by a classical application software and add it to institutional memory of the corporation [4] [5][6]. In this study, we offer an expert system model which will analyze all working process and data cluster of a corporation by using text mining method which works independent of open source software but using them as data source.

II. RELATED WORKS

Expert system model which uses text mining methods is a multidisciplinary working area containing information, data mining, machine learning, statistics, additional language science concepts inside. Beginning from 2000's, parallel to the technologic developments it started to be used on fields due to the needs of government institutes and business world [7]. A lot of studies are made suitable to these needs including national security and information applications, forgery detection, medical and health applications, marketing research solutions, emotional analyze tools, publishing applications, social media tracing and analyze, searching and knowledge access solutions, advertising applications, institutional work intelligence solutions [7][9][10][11][12].

III. MOSTLY USED OPEN SOURCE BUSINESS AND PROCESS MANAGEMENT SOFTWARE

Today, a corporation must be developing in work process continuously to raise its knowledge and experience. There are lots of open source work and process management software to be used for this. İt is required to select a suitable software for all data of corporation to develop institutional memory[9][10][13]. Here are the most important features which a work and process management software must have[14][15][16].

- To be able to collect the documents in a central area and associate them to institute, person or categories
- To be able to manage all work process of clients centrally
- To be able to save the activities which are fulfilled by corporation in the system (sales, telephone numbers, visits, meetings etc.)
- To save all documents about offers, commitments etc. in the system
- To save e-mail, fax and other mail's records
- To be able to associate all data input about work process
- To be able to create secure data saving of corporations and built the system for backup and authorizing.
- To be able to arrange, update and search all known file formats
- To be able to prepare reports which are required about corporation's documents

Commonly used open source work and process management software:

Pentaho Community Edition: Pentaho Community Edition OLAP is a complicated work process management solution which contains analyze, data mining, reporting tools and management interfaces. JasperReports Business Intelligence: JasperReports is an open source reporting library which has a strong architecture design and usable with different kinds of applications. It is possible to have reports in pdf,odt, xml, html, cvs, rtf etc. formats, using Jasper Reports in different areas from web to swing applications. JasperReports report design template enables to make a report template which is flexible and built of independent sectors. And it is possible to divide the report into sub reports and use them in different orders in the other reports. BIRT Business Intelligence and Reporting Tools: BIRT is an open source Eclipse based reporting tool. İt is used to create reports in Java applications. Jedox Base: there is no limitation about size, users, elements and number of properties in Jedox Base Business Intelligence solution. All data is kept in JedoxOLAP server. Jedox is able to make multi-dimensional and real time computing and processing.

ReportServer: ReportServer is a highly developed work intelligence and reporting application which can be used with a GPL license, with hierarchical structure, dynamic listing, graphically and advanced reporting options, interactive management panels, scheduled task definitions, advanced searching ability, and data management features[15][16].

IV. TEXT MINING METHODS AND OPEN SOURCE STUDIES

Text mining is to get the useful and unknown knowledge with various methods and technics from the data in text form. Text mining is a study area which aims to have the significant data within the big data by using the data mining, artificial intelligence, NLP Natural Language Processing, statistics, IR Information Retrieval and knowledge management technics. İt is possible to divide text mining technics into four main categories[14][15][16];

Classification: Classification process is to attach the objects to previously known classes or categories. Association analysis: Association analysis is the task of uncovering relationships among data, identifying how the data items are associated with each other.
Information extraction: It is trying to find useful data within the documents with information extraction technics. Clustering: Clustering analyses is applied to discover document structures By using this technics it is possible to make some studies like; classification of documents, clustering, concept/entity extraction, production of granular taxonomy, sentimental analysis, document summarization, entity relationship modelling. In this respect, as a part of text mining studies the methods like information retrieval, lexical analysis, Word frequency distribution, pattern recognition, tagging, information extraction, data mining and even visualization.

Commonly used open source text analyze and text mining applications:
QDA Miner Lite: QDA Miner is used for examining and analyzing text based data. It is possible to analyze the text files saved on Excel, MS Access, CSV or RTF, HTML, PDF documents. The software is also able to make fast search on documents and analyze the graphics in bar or pie chart.

Gate: GATE is a text based analyze application which is able to make natural language process. Gate is used in many areas like customer voice recognizing, cancer researches, medicine researches, decision support processes, employing process, web mining and data extraction.

TAMS Analyzer: ıt is an application works on Macintosh OS x. It is used to determine the themes analyzing web pages, interviews. Carrot2: carrot2 makes clustering on texts and search results. İt is an open source application which puts the clustering it makes into categories automatically. It is able to take and analyze the search results from GoogleAPI, Bing API, eTools Meta Search, Lucene, Solr and so on. CAT: CAT is a service of a quality data analyses programme. By this service raw text data sets are coded effectively. It can administrate the coding permissions of team on web service. It can open unlimited numbers of accounts for collaboration. It can assign more than one programmer for one work to work on it in a cooperation. KH Coder: KH Coder is an application advanced for qualified content analysis and text mining. It is possible to process language data in Japanese, English, French, Portuguese and Spanish.

It is possible to make search, statistical analyze, multidimensional scaling, cluster analyze on the raw text input. Gensim: gensim is a python library used to analyze text files. Rapid miner: rapid miner is an integrated context application for machine learning, data mining, predictive analytics and business analytics[14][15][16].

V. TEXT PROPOSED MODEL FOR INSTITUTIONAL MEMORY TO PRODUCE THE BIG DATA

Text mining is a topic which aims to extract significant information from big data source with using data mining, natural language processing, statistic, information retrieval, knowledge management techniques. It is a way to achieve the useful and previously unknown information from text formatted data with data processing methods and techniques. This study are mentioned on the contribution of text mining methods to improve the institutional memory [17][18][19][20][21].

Qualitative data should be passed through some process steps to run Data mining algorithms. Pre-processing, cleaning, conversion are some of these steps. Data should be transformed by Pre-processing steps. This step involves the provision of digital conversion of the data. Cleaning step should be used for data which is not necessary information sweeping. This step is followed by a subsequent process steps required to undergo classification of data. This step will perform the interpretation and data modeling steps. With the proposed model, providing qualitative data classification and clustering, making the extraction information is recommended.

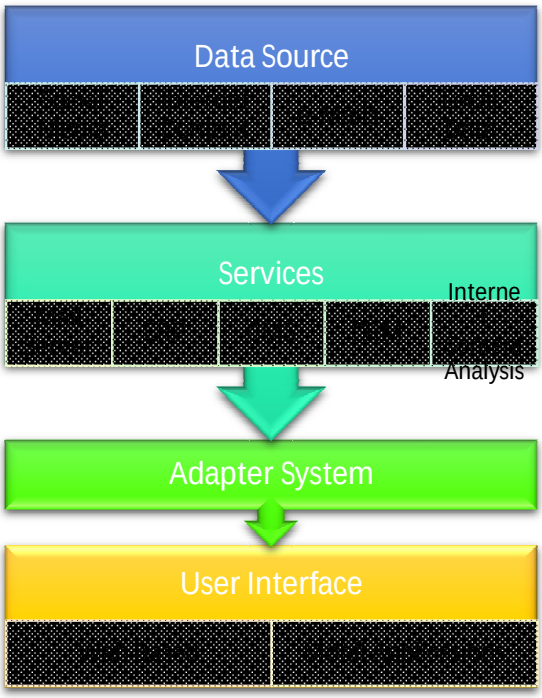

Figure 1The proposed model

Steps that will allow the conversion of meaningful information from big data in corporate base is given in the figure 1. Model should work with the flow of Social media, Internet Contents, emails. All services which is in need of an institution is accessible as open source. The proposed system is expected to run on the data which produced by the open source software according to corporate needs.

Data streams must be connected to the adapter system that will work with. Adapter structure should be used to evaluate the data obtained through the stack of open source systems. Data sources are independent of each other but have a critical relationship to the organization in terms of content. This shows how a revealed text mining methods with the processing efficiency of the data flowing.

Evaluation of the data which obtained from sources that produce multiple enterprise information with the proposed structure of adapter should be provided only an expert system. This will help to put forward a relational data structure which is independent of the service. Adaptor system which contains text mining and expert system algorithms will be able to reveal their relationship with each other qualitative data by more than one service. It is planned to produce useful information in this relationship with the corporate base. New generation of expert systems that will be proposed under study is planned to be a guidance document on integrated ERP software. The expert systems, the elimination of individual errors in the assessment and the process will be carried out very quickly and provides an easy way.

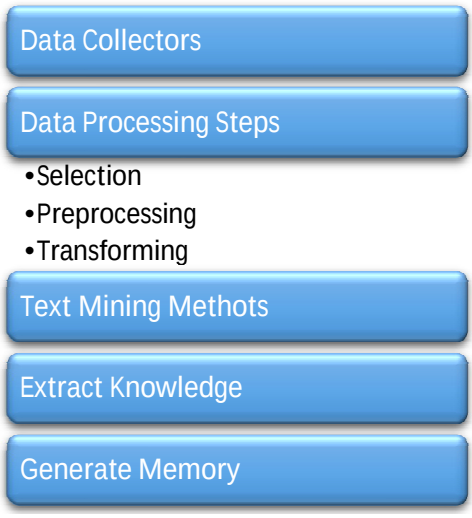

Figure 2Model adaptor

The task of adapting the system generates important corporate information or alarms. The information which affects the efficiency of the generated information should be given to user with interfaces which is given in figure 2 as final step. It is important that the interface of the web and mobile based. Established model includes an idea proposed the use

of open source enterprise information system to produce a text mining.

This is a decision to work with the model proposed by the institutional knowledge from raw data to obtain support system model reveals. Evaluation decision support system will help to increase the quality of information which is produced by local or external sources.

CONCLUSIONS

Text mining is an important place in the meaning of qualitative data. Especially if the majority of enterprise information is assumed to be on the texts, text mining algorithms used by corporate foundations in open source systems will be shown to increase the yield. The proposed expert system, future predictions about the data collected may be found in institutions and it can generate tag or topic based alarms with this estimations.

REFERENCES

[1] . Dan Sullivan, “Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales”, 2001, John Wiley & Sons, Inc. New York, NY, USA

[2] . Alla Keselman, Vimla L Patel, Todd R Johnson, Jiajie Zhang , “Institutional decision-making to select patient care devices: identifying venues to promote patient safety” , Journal of Biomedical Informatics, Volume 36, Issues 1–2, February–April 2003, Pages 31–44.

[3] . Serkan Altuntas, Turkay Dereli, Andrew Kusiak, “Analysis of patent documents with weighted association rules”, Technological Forecasting and Social Change, Volume 92, March 2015, Pages 249–262

[4] . Hanna Suominen, “Text mining and information analysis of health documents”, Artificial Intelligence in Medicine, Volume 61, Issue 3, July 2014, Pages 127–130

[5] . Heeyong Noh, Yeongran Jo, Sungjoo Lee, “Keyword selection and processing strategy for applying text mining to patent analysis”, Expert Systems with Applications, Volume 42, Issue 9, 1 June 2015, Pages 4348–4360

[6] . Michael A. McDaniel, Bryan J. Pesta, Allison S. Gabriel , “Big data and the well-being nexus: Tracking Google search activity by state IQ”, Intelligence, Volume 50, May–June 2015, Pages 21–29

[7] . Shaokun Fan, Raymond Y.K. Lau, J. Leon Zhao, “Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix”, Big Data Research, 2015.

[8] . Clemens Költringer, Astrid Dickinger, “Analyzing destination branding and image from online sources: A web content mining approach”, Journal of Business Research, 2015.

[9] . Yen-Liang Chen, Kwei Tang, Chia-Chi Wu, Ru-Yun Jheng, “Predicting the influence of users’ posted information for eWOM advertising in social networks”, Electronic Commerce Research and Applications, Volume 13, Issue 6, November–December 2014, Pages 431–439.

[10] . Donald R. Mendoza, Ronald Johnson, “Using a Lessons Learned Process to Develop and Maintain Institutional Memory and Intelligence”, Aerospace Conference, 2006 IE.

[11] . Ulas M., Tasci S.M., Ulas S., "Augmented Reality and Application Sample on Education" ,12/2014, Institute of Research Engineers and Doctors, USA, ISBN: 978-1-63248-034-7, doi: 10.15224/ 978-1-63248-034-7-47 , S.116-118

[12] . Aldim U.F., Ulas M., “The Use of Social Media Elements in Distance Learning”, Journal of Teaching and Education, ISSN: 2165-6266, 03(02):233–239 (2014).

[13] . Ulaş M. and Boyacı A., "Development of a Hierarchic Content Management System", e-journal of New World Sciences Academy, Engineering Sciences, ISSN:1308-7231, Vol.7, No.1. pp 1-13, 2012.

[14] . Access: 01.03.2015, http://www.predictiveanalyticstoday.com/bigdata-platforms-bigdata-analytics-software/

[15] . Access: 01.03.2015, http://www.predictiveanalyticstoday.com/top-30-software-for-text-analysis-text-mining-text-analytics/

[16] . Access: 01.03.2015, http://www.predictiveanalyticstoday.com/open-source-free-business-intelligence-solutions/

[17] . G.V. Asokana, Vanitha Asokan, “Leveraging “big data” to enhance the effectiveness of “one health” in an era of health informatics”, Journal of Epidemiology and Global Health, 2015

[18] . Ömer M. Soysal, “Association rule mining with mostly associated sequential patterns”, Expert Systems with Applications, Volume 42, Issue 5, 1 April 2015, Pages 2582–2592

[19] . Thanh Nguyen, Abbas Khosravi, Douglas Creighton, Saeid Nahavandi, “Classification of healthcare data using genetic fuzzy logic system and wavelets”, Expert Systems with Applications, Volume 42, Issue 4, March 2015, Pages 2184–2197.

[20] . Wilco W.M. Fleuren, Wynand Alkema, “Application of text mining in the biomedical domain”, Methods, Volume 74, 1 March 2015, Pages 97–106

[21] . Fernando Roda, Estanislao Musulin, “An ontology-based framework to support intelligent data analysis of sensor measurements”, Expert Systems with Applications, Volume 41, Issue 17, 1 December 2014, Pages 7914–7926.

★★★