Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing

F. S. Gharehchopogh
Hacettepe University Department
Computer Engineering
Ankara, Turkey

Z. A. Khalifelu
IAU Branch of Shabestar
Department of Computer Engineering
Shabestar, Iran

***Abstract*-Nowadays, most of information saved in companies are as unstructured models. Retrieval and extraction of the information is essential works and importance in semantic web areas. Many of these requirements will be depend on the storage efficiency and unstructured data analysis. Merrill Lynch recently estimated that more than 80% of all potentially useful business information is unstructured data. The large number and complexity of unstructured data opens up many new possibilities for the analyst. We analyze both structured and unstructured data individually and collectively. Text mining and natural language processing are two techniques with their methods for knowledge discovery form textual context in documents. In this study, text mining and natural language techniques will be illustrated. The aim of this work comparison and evaluation the similarities and differences between text mining and natural language processing for extraction useful information via suitable themselves methods.**

I. INTRODUCTION

In recent years, as a result of the development of information systems and technology, and businesses and other organizations databases, depending on the organization's purpose and structure of various types of data are collected [4]. However, raw data is not processed as long as a meaningless pile of data stored in databases [1]. Development of appropriate software and the company has collected the available data conversion request to information collected by this data processing, data can be used within interesting relationship and patterns appears to be making was required. Today's, many organization customer data and customers' purchasing patterns related to quality useful [4].

Useful information will be obtained with methods not begin processing. Raw data is rich, knowledge in the event of poor institutions that succeed in competitive markets and to continue their success even more difficult with passing day. Who understand the importance of data collection and query data with retroactive benefits from the highest of all institutions cannot begin to see the biggest help, is data mining. Meaningful information from raw data can be produced by data mining. Techniques used to extract relationships.

Data mining can analyze structured data, while the text and web mining unstructured data is used to analyze and transform data become structured [3, 5]. On the other hand, Natural Language Processing (NLP) applies unstructured data for finding and realizing natural language and textual information. In this study, we analyze both structured and unstructured data. Thus application of data, text, and web mining will be described. Then we explain and review of NLP and text mining usages in text processing. Finally, similarities and differences among of them have investigated.

II. DATA, TEXT AND WEB MINING

Data Mining is knowledge detection and resolution process of databases [3, 6]. It has obtained previously unknown, secret, meaningful and useful patterns being automatically established from large scaled databases [14, 15]. So, data mining knowledge discovery in databases is looking for patterns in data [10]. Likewise, text mining looking for patterns in text.

Text mining is the process of analyzing text to extract information that is useful for particular purposes. Text is unstructured, amorphous, and complicated to deal with. Nevertheless, text is the most common vehicle for the formal exchange of information.

Data mining algorithms contains a combination statistical algorithms, mathematical algorithms and artificial intelligence algorithms (such as neural networks, decision trees, cohune networks, union rules and etc) [10]. Generally Data mining can analyze structured data. Data mining tools and algorithms in text or web to be finding patterns in data or model before creating the text or web of data must be structured [4, 8 and 9]. As well as text mining could be considered information retrieval as text retrieval or document retrieval, what search engines do [16] and etc in text classification area. Text and Web mining operations

978-1-61284-832-7/11/$26.00 ©2011 IEEE

will be used in data mining tools to be access structured data can be described [2, 7].

Data mining techniques and their tools [18] are designed to exert structured data from databases. Text mining functionality [19] is similar to data mining, but text mining can work with unstructured data such as PDF files or semi-structured data sets such as emails, XML and HTML files and etc. So, text mining is a superior way for companies in business fields. Since the most of information in these places is saved as text or in text files [17].

III. Natural Language Processing

Natural Language Processing (NLP) from 1960 is developed as yet. This research area is sub filed of Artificial Intelligence (AI) and linguistics regions. Main aim of NLP studying are generation and realizing of natural languages. So, by means of these methods of NLP for processing amounts of textual information is considered very efficient and intelligent. One direction of NLP research relies on statistical techniques, typically involving the processing of words found in texts [20]. One of the NLP applications in text retrieval is usage of these techniques as a necessary component in web search engines, via automated translation tools or in summary generators [21].

By means of NLP techniques, new approach creates usage of rule based methods, leveraging knowledge resources. These can include items such as ontologies and linguistic rules. The statistical human language processing systems need sets of training principled as if indicate the desirable (and/or undesirable) relations and dependencies [22]. In Artificial Intelligence ontologies are developed by humans as models [26]. Ontology serves as a representation vocabulary that provides a set of terms with which to describe the facts in some domain [26].

NLP applications in wide area of AI consist of question answering systems [23], automatic translate of languages to each other, opinion mining systems and etc.

Some of the issues facing the NLP systems contain linguistic variation and ambiguity. The linguistic variation means the possibility of using different words or expressions to communicate the same idea while Linguistic ambiguity is when a word or phrase allows for more than one interpretation [21, 25]. In other hand, Ontology applications in NLP research area are including ontology provides a context for the vocabulary it contains [26, 27 and 28].

IV. Unstructured Data

Most previous studies of data mining have focused on structured data, such as relational, transactional, and data warehouse. However, a substantial portion of the available information is stored in text databases, which consist of large collections of documents from various sources, such as news articles, research papers, books, digital libraries, email messages, and Web pages.

Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web [4].

Nowadays most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases. Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured [7].

For example, a document may contain a few structured fields, such as title, authors, publication date, and category, and so on, but also contain some largely unstructured text components, such as abstract and contents [3]. There have been a great deal of studies on the modeling and implementation of semi structured data in recent database research. Moreover, information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents.

Traditional information retrieval techniques [11, 12] become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining [12].

Structured data, that can be organized structure and therefore can be defined a term used for the actual data. The most commonly used universal type of structured data such as SQL and Access are data sources. For example, Structured Query Language (SQL), columns (variables) and rows (records) based information allows in select. The content of structured data can be organized according to the data types and data is searchable [11].

Unstructured data refers to usually computerized information that either does not have a data model nor has one that is not easily usable by a computer program [8]. Unstructured data distinguishes such information from data stored in fielded form in databases or annotated in documents. Probably the most common types of unstructured data such as image files, PDF, word and text, are kept text files on the web and e-mail log files [11]. In spite of organize E-mail databases with tools such as Microsoft Outlook, which kind of structured data is consider to change raw data.

Structured data types have cell structure such as Excel, although whether or not to present the structured locations are discussed [13]. Many institutions provide unstructured data in databases. Merrill Lynch in 1998 cited estimates that as much as 80% of all potentially usable business

information originates in unstructured form [3]. Such estimates may not be based on primary research, but they are nonetheless widely accepted. This is important given that company's information must be classified. So, data can be classified in the following ways include data type, Data organization, data age and data value [26].

However, the problem of Knowledge Discovery from Text (KDT) [6] is to understand explicit and implicit concepts and semantic relations between concepts existing in unstructured files using NLP techniques [17].

V. EVALUATION

Text mining research area for finding patterns in texts includes search and retrieval, document navigation and exploration, text analysis, knowledge management, extraction of topics from texts or groups of text and the analysis of topic trends in text streams[16]. While information retrieval and other forms of text mining frequently make use of word stemming, more sophisticated techniques from NLP have been rarely used [16, 32].

. Text classification for region specific databases and path finding by means of reinforcement learning methods are applying for text patterns in this extent [31].

Search function based on two types of discoveries. First is based on content consist of clustering and text categorization. Second based on concept contain predictive modeling, associative discovery, deviation detection, trend analysis. Access methods for texts documents are direct browsing in web environment and information retrieval. Since, learning-based text categorization is the simplest form of text mining [25].

Text mining techniques primarily developed in the domains of information retrieval, statistics, and machine learning [31, 32]. Its aim typically is not to understand all or even a large part of what a given speaker/writer has said, but rather to extract patterns across a large number of documents [16]. They explore Naive Bayes and SVMs to perform the text classification, they also use grammatical roles derived from an NLP parser [16].

Generally, text mining is extraction of interesting and useful patterns in text data but NLP technologies is as information discovery and NLP techniques supply text classification, text categorization, document clustering, finding groups of similar documents, information extraction, summarization and etc. these are content based techniques. Text classification and bayesian networks are two popular algorithms in text mining for finding and matching patterns in texts. They explore Naïve Bayes and SVMs to perform the text classification, but they also use grammatical roles derived from an NLP parser [16].

Hence the problem of *where* and *how* to supply unstructured data in order to find it efficiently, needs a fresh re-assessment [24]. As for text mining and NLP in intelligent text classification, as a result both of text mining and NLP are two methods for processing textual information. This work is starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets [17] and information extraction.

Using NLP techniques, the entities and relations that act as indicators of recoverable claims are mined from management notes [26]. Text mining techniques can then be applied to find dependencies between different entities, and to combine indicators to provide scores to individual claims. [22].

NLP techniques are used for text that is typically syntactically parsed using information from a formal grammar and a lexicon, the resulting information is then interpreted semantically and used to extract information about what was said [16].

NLP includes techniques like word stemming (removing suffixes) or a related technique, lemmatization (replacing an inflected word with its base form), multiword phrase grouping, synonym normalization, part-of-speech (POS) tagging (such as elaborations on noun, verb, preposition and etc), word-sense disambiguation, anaphora resolution and role determination (such as subject and object)[16].

We will explain the functions and peculiarities of the two key approaches to natural language processing: a statistical approach and a linguistic focus. Statistical processing of natural language [29] represents the classical model of information retrieval systems, and is characterised from each document's set of key words, known as the terms index.

In NLP document processing for knowledge discovery consist of document pre-processing and Parameterisation. This approach is based on the application of different techniques and rules that explicitly encode linguistic knowledge [30].

In document pre-processing level fundamentally consisting in preparing the documents for its parameterisation, eliminating any elements considered as superfluous and in parameterisation level is a stage of minimal complexity once the relevant terms have been identified[16].

With NLP techniques, the documents are analysed through different linguistic levels by linguistic tools that incorporate each level's own annotations to the text [29]. After having identified and analysed the words in a text, the next step is to see how they are related and used together in making larger grammatical units, phrases and sentences [22, 24]. The techniques used to apply and create parsers vary and depend on the aim of the syntax analysis.

VI. CONCLUSION AND FUTURE WORKS

Structural data obtained using the unstructured model of the data using methods of text and web mining be brought, and from there obtained structural models were compared. The results obtained, text and web mining methods using the obtained model is more successful shows. Unstructured

data model to integrate the quality of information can be removed and this result also shows that the unexpected results. Potentially used in the world 80% of all unstructured types of data considered, the use of this data will certainly add value to research. The other important issue is evaluation unstructured data analysis in text mining methods and NLP techniques. Text mining try to finding patterns in textual unstructured files based on contents. NLP try to reach concepts of texts via specific algorithms.

REFRENCES

[1] H. Jiawei and K. Micheline, 2006, |Data Mining: Concepts and Techniques|, *vol. 2*, Morgan Kaufmann Publisher.

[2] F. Ronen and S. James, 2006, |The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", *vol. 1*, Cambridge University press.

[3] F.S. Gharehchopogh, |Approch and Review of User Oriented Interactive Data Mining", *IEEE, the 4th International Conference on Application of Information and Communication Technologies(AICT2010),* Uzbekistan, pp. 1-4, 2010.

[4] K. Manu, 2006, "Text Application Programming (Programming Series)", *vol. 1*, Chartless River Media.

[5] F. Ronen and S. James, 2006, |The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", *vol. 1*, Cambridge University press.

[6] F. S. Gharehchopogh, "Approach and Developing Data Mining Method for Spatial Applications", *Proceedings of International Conference on Intelligent Systems & Data Processing (ICISD),* India, 2011, pp. 342-345.

[7] W. Berry Michael, 2006, "Survey of Text Mining: Clustering, Classification and Retrieval", *vol. 2*, Springer.

[8] Y. Shiqun, Q. Yuhui, G. Jike and W. Fang,2008, "A Chinese Text Classification Approach Based on Semantic Web", *Fourth International Conference on Semantics Knowledge and Grid*, pp. 497-498.

[9] L. Rui and J. Minghu, 2008, "Chinese Text Classification Based on the BVB Model", *2008 IEEE, Fourth International Conference on Semantics Knowledge and Grid*, pp. 376-379.

[10] L. Huo, F. Yi and H. Heping, 2008, "Dynamic Service Replica on Distributed Data Mining", *2008 IEEE International Conference on Computer Science and Software Engineering*, PP. 390-393.

[11] Y. Shiqun, W. Gang, Q. Yuhui and Z. Weiqun, "Research and Implement of Classification Algorithm on Web Text Mining", *2007 IEEE, Third International Conference on Semantics Knowledge and Grid*, 2007, pp. 446-449.

[12] W. Fan, W. Linda, S. Rich and Z. Zhang, "Tapping into the power of text mining", 2005 USA, *Communications of the ACM*, pp.76-82.

[13] A. H. Tan, P. S. Yu, 2004, "Guest Editorial: Text and Web mining", 2004 USA, *Applied Intelligence*, pp. 239-241.

[14] H. Witten Jan, F. Eibe, "Data Mining: Practical Machine Learning Tools and Techniques". Diane Cerra Publishers, 2005.

[15] Weiss S, Indurkhya N, Zhang T, Damerau F, "Text Mining: Predictive Methods for Analyzing Unstructured Information". *Springer*, 2004.

[16] A. Kao, S. Poteet, "Text Mining and Natural Language Processing-Introduction for the Special Issue", *SIGKDD Explorations, 2004, vol. 7*, Issue. 1, pp. 1-3.

[17] V. Gupta, "A Survey of Text Mining Techniques and Applications", *Journal of Emerging Technologies in Web Intelligence, vol. 1, No. 1,* 2009, pp. 60-76.

[18] Navathe, Shamkant B., and Elmasri Ramez, (2000), "Data Warehousing And Data Mining", in "*Fundamentals of Database Systems*", Pearson Education pvt Inc, Singapore, 841-872.

[19] Berry Michael W., "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", *Springer Verlag*, USA, LLC, 2004, pp. 24-43.

[20] Manning, C., Schutze, H., "Foundations of Statistical Natural Language Processing", MIT Press, Cambridge, MA, 1999.

[21] Baeza Yates, "Challenges in the Interaction of Information Retrieval and Natural Language Processing", *in Proceedings of 5th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Corea, Lecture Notes in Computer Science, Springer, 2004, vol.* 2945, pp. 445-456.

[22] F. Popowich, "Using Text Mining and Natural Language Processing for Health Care Claims Processing*", SIGKDD Explorations*, vol. 7, Issue. 1, pp. 59-66.

[23] N. Pala Er, "Turkish Factoid Question Answering Using Answer Pattern Matching", MS.c Thesis, Bilkent University, 2009.

[24] A. Raghuveer, M. Jindal, M. F. Mobel, B. Debnath, D. Du, "Towards Efficient Search on Unstructured Data", *An Intelligent-Storage Approach, CIKM'07, ACM*, Portugal, 2007, pp.951-954.

[25] D. Bhattacharyya, S. Biswas, T. H. Kim, "A Review on Natural Language Processing in Opinion Mining", *International journal of Smart Home, vol. 4, no. 2*, 2010, pp. 31-38.

[26] A. K. Goel, R. Sindhu, M. Mehrotra, G. N. Purohit, "Managing Unstructured Data Using Agent Technology", *UbiCC Journal, vol. 45, no. 3*, 2009, pp. 801-806.

[27] Albers M, Jonker CM, Karami M,Treur J., "Agents models and different user ontology's for an electronic market place", *Knowl Inf Syst, 2004, vol. 6, no. 1*, pp. 1-41.

[28] Alexander Smirnov & Nikolay Shilov, "Ontology-driven intelligent service for configuration support in networked organization", *Springer-Verlag*, London Limited, 2007.

[29] Manning, C. D. and Schütze, H. "Foundations of statistical natural language processing", MIT Press. Cambridge, MA: May 1999, p. 680.

[30] Sanderson M., Retrieving with good sense, "In: Information Retrieval", *2000, vol. 2,* pp. 49-69.

[31] *KRDL's Text Mining* site available at: http://textmining.krdl.org.sg/resources.html, last access 29/6/2011.

[32] *KDNuggets: Data Mining and Knowledge Discovery Resources* available at: http://www.KDNuggets.com, last access 29/6/2011.