Structured and Unstructured Information Extraction Using Text Mining and Natural Language Processing Techniques

S. Nagarajan
Department of Computer Applications,
School of Information Technology,
Madurai Kamaraj University,
Madurai, Tamilnadu,India
nagasethu2000@yahoo.com

Dr. K. Perumal
Department of Computer Applications,
School of Information Technology,
Madurai Kamaraj University,
Madurai, Tamilnadu,India
perumalmkucs@gmail.com

***Abstract*—-** Information on web is increasing at infinitum. Thus, web has become an unstructured global area where information even if available, cannot be directly used for desired applications. One is often faced with an information overload and demands for some automated help. Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents by means of Text Mining and Natural Language Processing (NLP) techniques. Extracted structured information can be used for variety of enterprise or personal level task of varying complexity. The Information Extraction (IE) in also a set of knowledge in order to answer to user consultations using natural language. The system is based on a Fuzzy Logic engine, which takes advantage of its flexibility for managing sets of accumulated knowledge. These sets may be built in hierarchic levels by a tree structure. Information extraction is structured data or knowledge from unstructured text by identifying references to named entities as well as stated relationships between such entities. Data mining research assumes that the information to be "mined" is already in the form of a relational database. IE can serve an important technology for text mining. The knowledge discovered is expressed directly in the documents to be mined, then IE alone can serve as an effective approach to text mining. However, if the documents contain concrete data in unstructured form rather than abstract knowledge, it may be useful to first use IE to transform the unstructured data in the document corpus into a structured database, and then use traditional data mining tools to identify abstract patterns in this extracted data. We propose a novel method for text mining with natural language processing techniques to extract the information from data base with efficient way, where the extraction time and accuracy is measured and plotted with simulation. Where the attributes of entities and relationship entities from structured and semi structured information .Results are compared with conventional methods.

Keywords: *Information Extraction (IE), Unstructured, semi structured, Data Mining, Natural Language Processing (NLP), Text mining (TM)*

*****

I. Introduction

The huge amount of documents on the web (or specifically, the web pages) by searching through a search engine or browsing through hyperlinks existed within web pages. Users which have no specific target often choose browsing web pages to achieve their final goal. However, many users have difficulty of getting start from a page which will eventually lead to their goals. Hence many portal sites emerge to provide such starting points. These sites often provide some sorts of navigating structure such as web directories or web hierarchies. Users can achieve a thematic navigation through such structures. However, these structures are generally constructed by human experts by hands, causing them lack of coverage and hard to maintain.

Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as title, authors, publication date, length, category, and, so on, but also contain some largely unstructured text components, such as abstract and contents. There have been a great deal of studies on the modeling and implementation of semi structured data in recent database research.[6] Information Retrieval techniques, such as text indexing, have been developed to handle unstructured documents. But, traditional Information Retrieval techniques 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 or 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.

I. Table Show the comparison between structured and unstructured data

Parameter	Unstructured Data	Semi Structured Data
Technology	Character and binary data	Relational database tables
Transaction Management	No transaction management and no concurrency	Matured transaction management, various concurrency techniques
Version Management	Versioned as a whole	Versioning over tuples, rows, tables etc.
Flexibility	Very flexible, absence of schema	Schema dependent, rigorous schema
Scalability	Very scalable	Scaling DB Schema is different

The above table shows that comparison between unstructured and semi structured data with respect to various parameters like technology, transaction management, version management, flexibility and scalability. In Technology basis, semi structure data having relational data base tables shown in results. In flexibility basis, rigorous schema possible than unstructured data.

- **Natural Language Processing (NLP)**

NLP is one of the oldest and most challenging problems in the field of artificial intelligence. The role of NLP in text mining is to deliver the system in the information extraction phase as an input[15].

- **Information Extraction (IE)**

Information Extraction[13] is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity includes processing human language texts by means of natural language processing (NLP)[14]. The recent activities in multimedia document processing like automatic annotation and mining information out of text images/audio/video could be seen as information extraction and the best practical and live example of IE is Google Search Engine.

Tasks performed by IE systems include:

Term analysis, which identifies the terms appearing in a document. This is especially useful for documents that contain many complex multi-word terms, such as scientific research papers.

Named-entity recognition, which identifies the names appearing in a document, such as names of people or organizations. Some systems are also able to recognize dates and expressions of time, quantities and associated units, percentages, and so on.

Fact extraction, which identifies and extracts complex facts from documents. Such facts could be relationships between entities or events.

Information Extraction=Segmentation + Classification + Clustering + Association

The information extraction process follow is described in this example

For years, Microsoft Corporation CEO Bill Gates railed against the Economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open- source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.
"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“ Richard Stallman, founder of the Free Software Foundation, countered saying...

The named entity recognition is first part of information extraction. It is called as segmentation.

Microsoft Corporation CEO Bill Gates Microsoft Gates Microsoft Bill Veghte Microsoft VP Richard Stallman Founder Free Software Foundation

II.TABLE RESULT OF INFORMATION EXTRACTION (example)

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	Founder	Free Software Foundation

- **Text**
- **Mining**

Text mining is the analysis of data contained in natural language text[16]. The application of text mining techniques to solve business problems is called *text analytics*. Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Face book, Twitter and LinkedIn. Mining unstructured data with natural language processing (NLP), statistical modeling and machine learning techniques can be challenging, however, because natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics, including slang, language specific to vertical industries and age groups, double entendres and sarcasm. The figure 1 is overview about the information extraction based text mining structure.

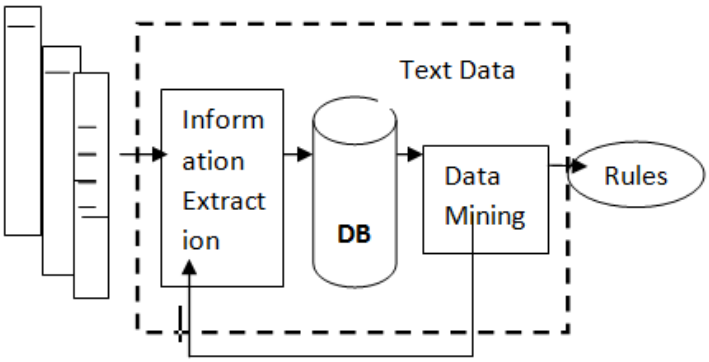

Fig. 1 Overview of IE based Text Mining Framework

- **Structured And Unstructured Text Mining Data**

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 data base 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 [20]. 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 [3]. For example, a document may contain a few structured fields, such as title, authors, publication date, and category, and so on.

Moreover, information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents. Traditional information retrieval techniques [21, 8] 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.

 ISSN: 2321-8169
32 – 43

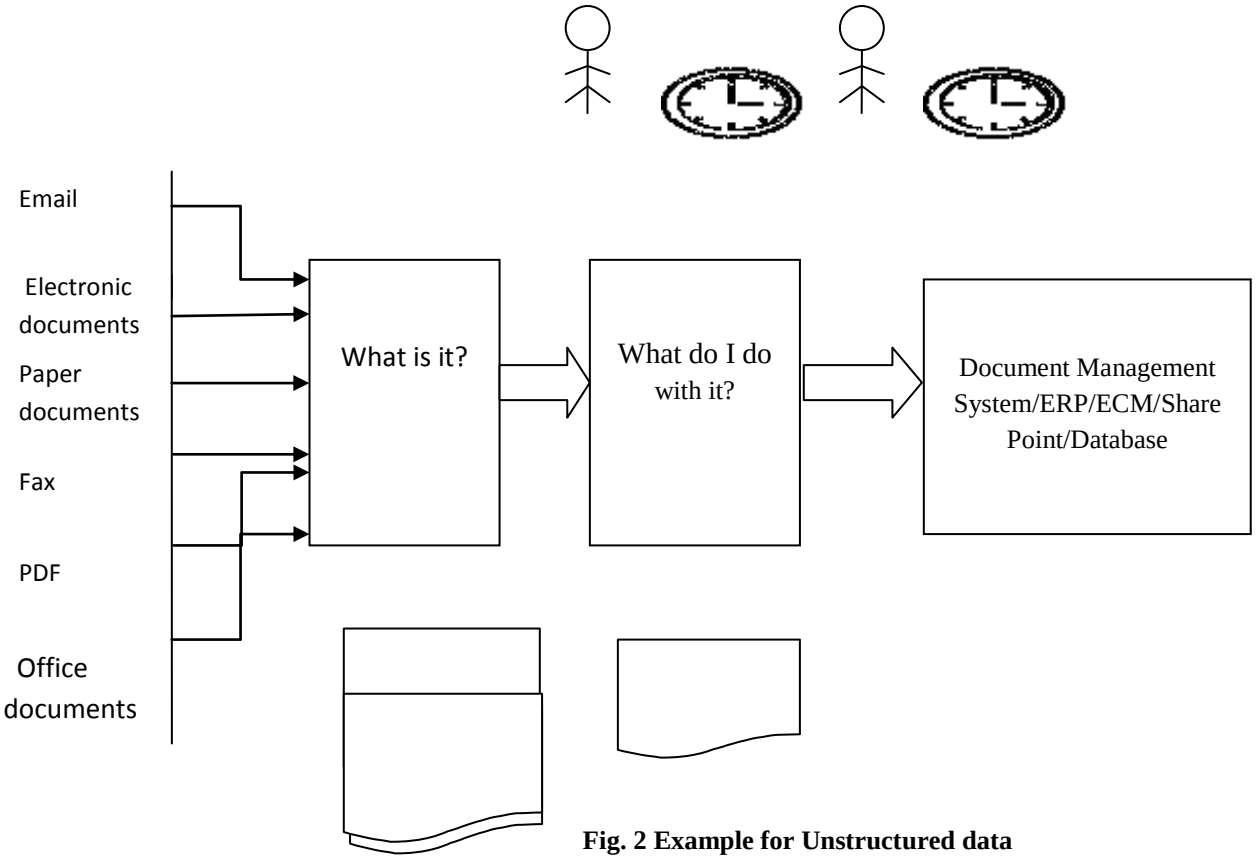

Fig. 2 Example for Unstructured data

Unstructured data [Fig. 2] refers to the information that no identical structure within this kind of data is offered. It's described as data, which cannot be kept in rows and columns in a relational database. Example for unstructured data is documented that is archived in folder, video and images. Structured Data [23] [Figure 2] refers predefined schema. The schema is instance information that conforms to this specification. The Example of structured data is a relational database system. Figure 3 illustrates an Entity Relationship diagram (ER diagram), its concrete on tables within an RDBMS [relational database management system].

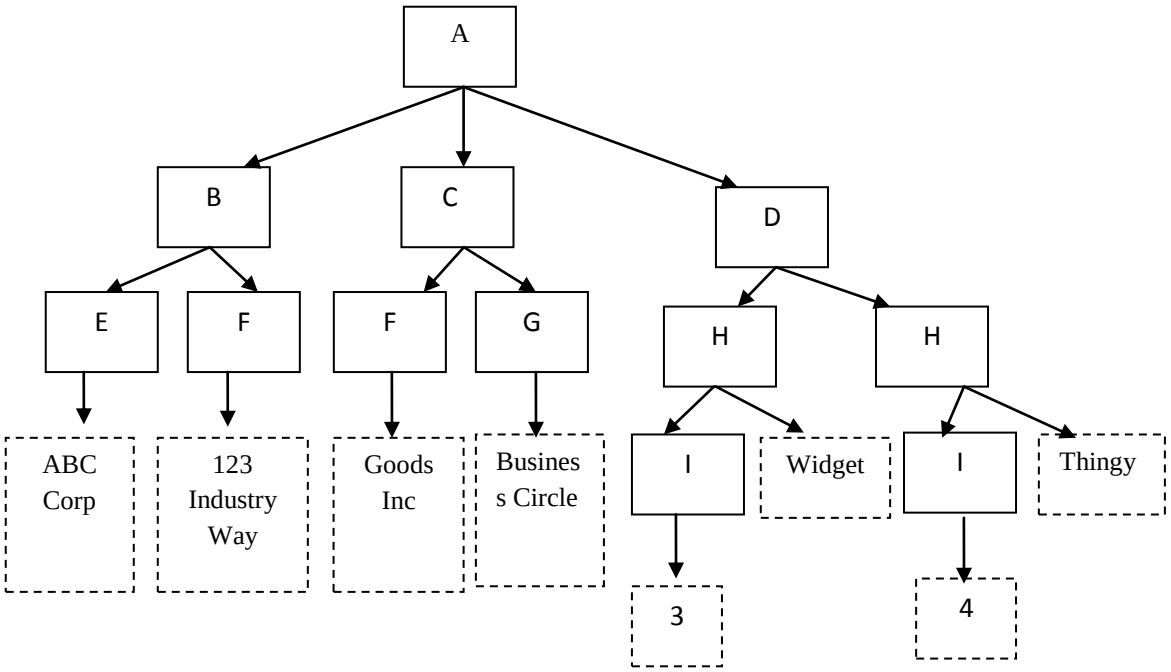

Fig. 3 Example for Structured data

- **Related Work**

The navigational structure in this work is a set of automatically identified category themes. Approaches on automatically generate the category themes are similar in context with research on topic identification or theme generation of text documents. Salton and Singhal [19] generated a text relationship map among text excerpts and recognized all groups of three mutually related text excerpts. A merging process is applied iteratively to these groups to finally obtain the theme (or a set of themes) of all text excerpts. Clifton and Cooley [7] used traditional data mining techniques to identify topic in a text corpus. They used a hyper graph partitioning scheme to cluster frequent item sets. The topic is represented as a set of named entities of the corresponding cluster. Ponte and Croft [18] applied dynamic programming techniques to segment text into relatively small segments. These segments can then be used for topic identification. Lin [12] used a knowledge based concept counting paradigm to identify topic through the Word Net hierarchy. Hearst and Plaunt [9] argued that the advent of full-length documents should be accompanied by the need for subtopic identification. They developed techniques for detecting subtopics and performed experiments using sequences of locally concentrated discussions rather than full-length documents. All these work, in some extent, may identify topics of documents that can be used as category themes for text categorization. However, they either rely on predefined category hierarchy (e.g. [12]) or do not reveal the hierarchy at all.

The application of intelligent agents in two aspects: the Knowledge Acquisition (KA) approach and the Information Extraction (IE) approach. Additionally, we produce a brief review of the existing studies related with knowledge extraction from natural language text documents. The techniques used in these studies involve knowledge acquisition agents, fuzzy logic intelligent agents, knowledge management, term weighting, databases knowledge discovery and biological text mining. The proposed work is mainly related to two areas of research: knowledge extraction from natural language text documents and knowledge modeling using intelligent agents.

Intelligent agent in KA and IE Zhiping, Tianwei, & Yu, (2010) produced a formal model of agent-based knowledge management in intelligent tutoring systems. It consists of three agents working together to construct, distribute, and maintain knowledge. The first one is a knowledge acquisition agent which is responsible for the construction of user model and domain knowledge base. The second is a knowledge distribution agent which is responsible for producing personalized teaching web pages to students dynamically. Finally, a knowledge maintaining agent which is responsible for the refinement of student models and domain knowledge[19].

Ropero, Gomes, Carrasco, & Leon, (2012) proposed a novel method for information extraction (IE) in a set of knowledge in order to answer to user consultations using natural language based on fuzzy logic engine. The sets of accumulated knowledge may be built in hierarchic levels by a tree structure. The aim of this system is to design and implementation a fuzzy logic intelligent agent to manage any set of knowledge where information is abundant, ambiguous, or imprecise. This novel method was applied to the case of university of Seville web portal which contain vast amount of information, they also proposed a novel method for Term Weighting (TW) based on fuzzy logic instead of using traditional term weighting scheme (TF-IDF)[11].

Knowledge extraction from text documents

Valencia-Garcia, Ruiz-Sanchez, Vivancos-Vicente, Fernandez-Breis, & Martinez-Bejar, (2004) produced an incremental approach for discovering medical knowledge from texts. The system has been used to extract clinical knowledge from texts concerning oncology. The authors started from notion of there are huge amounts of medical knowledge reside within text documents, so that the automatic extraction of that knowledge would certainly be beneficial for clinical activities. A user-centered approach for the incremental extraction of knowledge from text which is based on both knowledge technologies and natural language processing techniques is presented in this work. In same time, ontology is used to provide a formal, structured, reusable and shared knowledge representation [22].

- **Problem Definition**

To extract the rules, the IE task takes the set of tagged documents and produces a template representation for every document. This can be easily converted into rule-like form. For this purpose, a set of domain-independent extraction patterns are written so that we could match them against the input documents each extraction pattern constructs an output representation that involves two levels of linguistic knowledge: the rhetorical information expressed in the abstract and the semantic information contained in it, which we later convert into a predicate-like form. The left-hand expression states the pattern to be identified and the right hand side (following the colon) states the corresponding semantic action to be produced. A possible solution is to design a new IE methodology, subsequently referred to as TEMsIE1, which is based on various statistical and machine learning methods and techniques, selected to meet the requirements of the different IE phases. Although IE and data mining are similar in their general processing steps, they achieve different but complementary benefits. Consequently, augmenting IE with data mining creates a synergy that exploits benefits of both in order to improve the entire methodology significantly, especially its quality and performance. Time period is one most challenges in data extraction from data base. Here we give solution of time response of mining system with dimensional attributes.

- **Problem Solution**

Data mining research assumes that the information to be "mined" is already in the form of a relational database. IE can serve an important technology for text mining. The knowledge discovered is expressed directly in the documents to be mined, then IE alone can serve as an effective approach to text mining. However, if the documents contain concrete data in unstructured form rather than abstract knowledge, it may be useful to first use IE to transform the unstructured data in the document corpus into a structured database, and then use traditional data mining tools to identify abstract patterns in this extracted data. We propose a text mining with natural language processing technique e for extract the information from data base. We give novel model for extract the information from data base with dimensional attributes and retrieve with less timing for data extraction. Also the function of an expert mining technique is to convert unstructured human expertise to structured knowledge in a knowledge base to deal with easily. Intelligent agents or intelligent software agents use artificial intelligence in the pursuit of our objectives. We implemented Support Vector Machine (SVM) to check the text grammar in expertise extraction process.

- **Proposed Implementation**

The development of an intelligent agent that is capable to interview with a domain expert using question and answer in natural language, extract relevant knowledge from those answers, and convert these knowledge to a set of antecedence-consequence rules. At same time this agent also extracting a set of patterns or linguistic expressions and stores them into a conceptual database which is used later. The other intelligent agent is proposed in this work is to present a technique for extracting knowledge from natural language text documents. This agent tries to categorize input text documents into relevant or non-relevant documents with reference to a particular domain by calculating a threshold value based on term frequency-inverse document frequency (TF-IDF) for each input text document, Term Frequency (TF) is the frequency of occurrence of a term in a document and Inverse Document Frequency (IDF) varies inversely with the number of documents to which the term is assigned. TMIA tried to calculate the weight of term frequency for each feature (term) in text documents. Typically, the weights of term frequency for the phrase level are higher than for the word level because each element of the phrase level set consists of more than one word while the elements of the word level set consist of only one word. Eq(1) is the classical formula with some modification of TF-IDF used for term weighting:

$$TF - IDF(t_i, d_j) = \left(count(t_i, d_j) \times log \frac{|corpus|}{count - doc(t_i, corpus)} \right) \times r_i \qquad \dots (1)$$

where count(ti,dj) refers to the frequency of term ti in document dj, |corpus| refers to the number of documents in the corpus, count-doc(ti,corpus) is the number of documents in the corpus that contain the term ti, ri is the number of words in term ti which matched words in text documents. Eq(2) is the formula which calculates the average of all term weights calculated in eq(1):

$$Av - TF = \frac{\sum_{k=1}^{N} (TF - IDF(t_i, d_j))_k}{S} \qquad \dots (2)$$

Where TF-IDF(t_i,d_j) the weight of term frequency for the ith term, N is the number of features (terms), S is the total number of words that compose N. Now the TMIA should makea decision to accept or reject those text documents by compare the Av-TF value which is calculated in eq(2) with the threshold value as:

IF Av-TF >= Threshold value THEN accept OTHERWISE reject (3)

To accept means that, the produced text documents for processing are relevant to a particular problem diagnosis domain. In other words, it contains some relevant knowledge about that domain which can be extracted by other components of the proposed TMIA. On the other hand, to reject means that they are non-relevant to that domain and TMIA will ask the user to enter a new text documents file.

- **Hybrid Intelligent Mining Agent (HIMA)**

Expert mining is a process of extracting useful knowledge or meaningful patterns from human domain experts directly without interference of knowledge engineers, or it is a knowledge detection and resolution process of human domain experts. So, expert mining knowledge discovery in human brains which is looking for patterns of expertise looks like text mining which searches for patterns in text. Expert mining is a process of analyzing human expertise to extract knowledge (facts and rules) which is useful to solve particular problems in a specific domain. Expertise is unstructured or semi-structured, unorganized, and complicated to deal with; in other words, the function of an expert mining technique is to convert unstructured human expertise to structured knowledge in a knowledge base to deal with easily. Intelligent agents or intelligent software agents use artificial intelligence in the pursuit of their objectives.

The most important design considerations of intelligent agent systems is the agent interface as expert friendly as possible and hide the complexity of other components of the proposed agent. The intelligent agent is determined by the nature of its agent interface, as it is the part of the intelligent agent that interacts with the domain

expert using questions and answers in natural language and menu driven techniques to manage the interview between the domain expert and the HIMA to achieve the main goal of that agent. The fig.4 represents the intelligent agents important in optimize knowledge management. The data mining agents perform various functions of data mining. It is increasingly significant to develop methods and techniques. The knowledge discovery processes can be assisted by agents in order to increase the quality of knowledge and to processes the patterns. The agent plays main role in making decision processes in knowledge and text mining process.

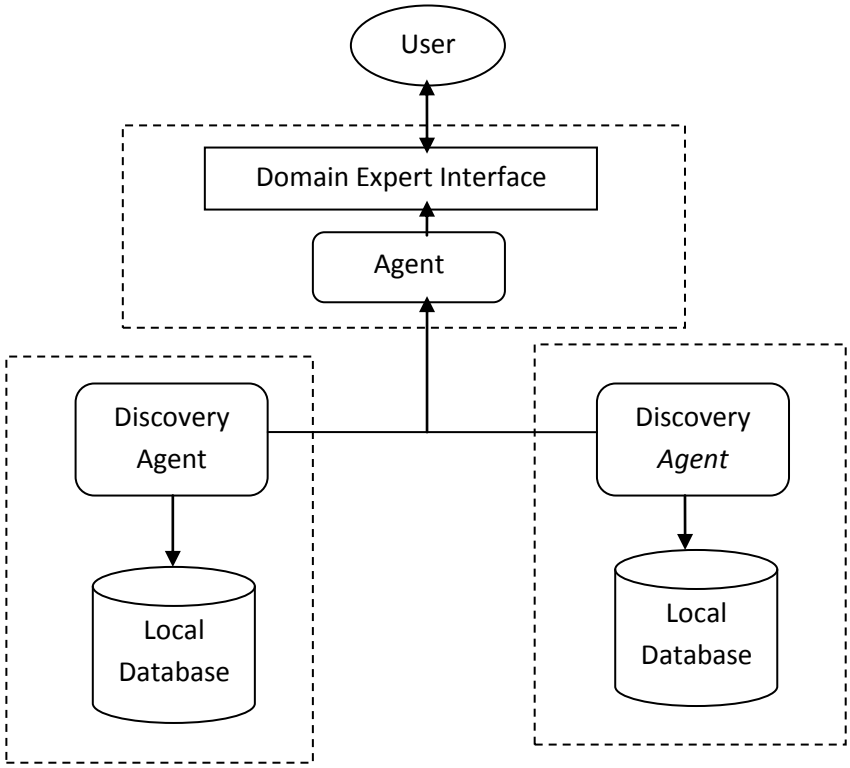

Fig. 4 Hybrid Mining with Agent Interface System

- **Knowledge Extraction**

The knowledge extraction process involves trying to extract knowledge from the interview with the domain expert by taking the expert's answers and putting them in a pre-formatted template which is prepared for that purpose.
This template consists of three parts as
1. Situation: it includes a main situation and a set of symptoms in the case of a production rule description or only a main situation in the case of a fact description.
2. Description: it includes all real descriptions for each situation in the first component as a sentence.
3. Pattern: it consists of two types of patterns; main pattern which represent the head of the rule (consequence) or fact, and sub-patterns which represent the body of the rule (antecedence). The difference between them is the first one has arguments and the other doesn't.

The card-pyramid parsing described in the previous section requires classifiers for each of the entity and relation productions. We use a Support Vector Machine (SVM) [4] classifier for each of the entity productions in the grammar. An entity classifier gets as input a sentence and a candidate entity indicated by the range of the indices of its words. It outputs the probability that the candidate entity is of the respective entity type. Probabilities for the SVM outputs are computed using the method by [5]. We use all possible word subsequences of the candidate entity words as implicit features using a word-subsequence kernel [2]. In addition, we use the following standard entity extraction features: the Part of Speech (POS) tag sequence of the candidate entity words, two words before and after the candidate entity and their POS tags, whether any or all candidate entity words are capitalized whether any or all words are found in a list of entity names, whether any word has “suffixment” or “ing”, and finally the alphanumeric pattern of characters [1] of the last candidate entity word obtained by replacing each character by its character type (lowercase, uppercase or numeric) and collapsing any consecutive repetition (for example, the alphanumeric pattern for CoNLL2010 will be AaA0). The full kernel is computed by adding the word-subsequence kernel and the dot-product of all these features, exploiting the convolution property of kernels. We also use an SVM classifier for each of the relation productions in the grammar which outputs the probability that the relation holds between the two entities. Parsing requires specifying a grammar for the card-pyramid. The productions in the

grammar are of two types. For leaf nodes, the productions are of the form entity Label, which stands for candidate entity, is the only terminal. The figure 5 [5]Show the knowledge extraction with agent intelligent systems connectivity. It provides the capability of dynamically incorporating knowledge. This knowledge has been extracted with the use of DM techniques.

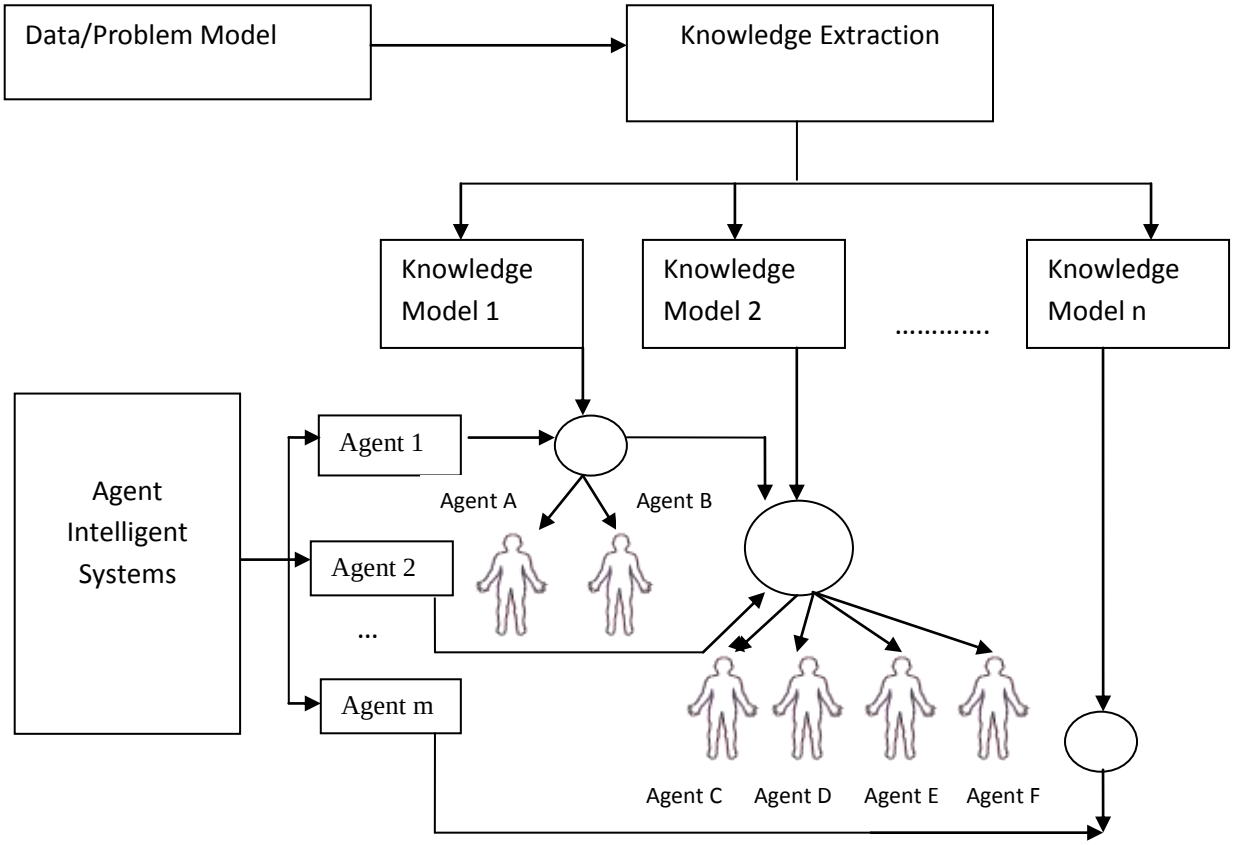

Fig. 5 Agent Intelligent System with Knowledge Extraction

Hybrid Intelligent Mining Agent (HIMA)

Expert mining is a process of extracting useful knowledge or meaningful patterns from human domain experts directly without interference of knowledge engineers, or it is a knowledge detection and resolution process of human domain experts. So, expert mining knowledge discovery in human brains which is looking for patterns of expertise, looks like text mining which searches for patterns in text. Expert mining is a process of analyzing human expertise to extract knowledge (facts and rules) which is useful to solve particular problems in a specific domain. Expertise is unstructured or semi-structured, unorganized, and complicated to deal with; in other words, the function of an expert mining technique is to convert unstructured human expertise to structured knowledge in a knowledge base to deal with easily. Intelligent agents or intelligent software agents use artificial intelligence in the pursuit of their objectives. They are capable of performing autonomous action in environments to achieve their goals . Wooldridge describes agents as computing entities which have four features: reactivity, autonomy, interaction, and initiative. Reactivity means that a system maintains continuous interaction with its environment if any change occurs in that environment. Autonomy is the main characteristic of agents. In other words, agents can cooperate autonomously to achieve predefined goal. Another aspect of agents is their ability to interact with other agents or humans using agent-communication language. Agents cooperating with each other can contribute to achieve goals because some goals can only be performed through cooperative work. Finally, the initiative feature of agents means they generate and attempt to perform goals by taking initiative instead of only being based on external environment events. These features of an intelligent agent clearly appear in the proposed HIMA. The reactivity feature is visible when the HIMA can work in different diagnosis domain; In other words, HIMA continues to work even if human the domain experts change. HIMA is capable a working independently to fulfill goals of the proposed agent and it also can cooperate, interact and contribute with other agents (TMIA) or humans (domain experts) to achieve the main goals of the multi-intelligent

agent (knowledge extraction). The HIMA also has the initiative feature through producing some questions to the expert domain and it can extract knowledge and linguistic expressions from experts' answers to save them in the knowledge base and a conceptual database respectively. Now let us discuss the components of proposed HIMA is explained above chapters.

III. Table Template for production using HIMA approach

Situation	Pattern	Description
Main situation	Disease(Slipped Disc)	Slipped Disc
Symptom1	Condition1	low back pain in the button of the back bone
Symptom2	Condition2	leg pain in left or right one
Main situation	Treatment("slipped disc", "Take drugs like Tilcotil20mg, if the situation continue do surgery operation").	Take drugs like Tilcotil20mg, if the situation continue do surgery operation

//*Pseudo Code for propose algorithm*//

```
disease_ name("slipped disc"):-
 condition1,
 condition2,
condition3, condition4.
conditions(condition1," low back pain in the button of the back bone").
conditions(condition2," leg pain in left or right one").
conditions(condition3," tingling, weakness, and foot senseless").
conditions(condition4," little legs narcotize or numbness").
treatment("slipped disc", "Take drugs like Tilcotil20mg, if the situation continue do surgery operation")
```

Results and Discussions

The assessment process of the proposed system is approved out across a medical domain. To be more precise, the referred domain is back pain diseases as in the above case study. For evaluating the show of the proposed system we can divide the process into 2 parts: HIMA evaluation and TMIA (Text Mining Intelligent Agent) evaluation.

$$\text{precision} = \frac{\text{number of documents correctly classified as positive by the system}}{\text{number of all documents classified as positive by the system}}$$

$$\text{recall} = \frac{\text{number of documents correctly classified as positive by the system}}{\text{number of positive documents in the testing set}}$$

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{accuracy} = \frac{\text{number of documents correctly classified by the system}}{\text{number of all documents in the testing set}}$$

In experimental setting environment, results are obtained from java environment with help of eclipse software with DB access. Each of the conventional and propose approaches was evaluate using cross-validation, a widely-used evaluation methodology for machine learning and text classification systems. A 95-fold cross validation was adopted, in which the 1500 ID in the data set were divided into 50 equal portions, with 40 documents each. Testing was performed for 100 iterations, in each of which 58 portions of the data (980 documents) were used for training and the remain portion (30documents) was used for testing. The data were rotated during the process such that each portion was used for testing in exactly one iteration.

In the conventional way, the knowledge engineer interviews domain experts to elicit problem solving information and formulates it in the information base. In the proposed architecture the HIMA interacts directly with area experts to extract their information and save it in the information base. We displayed HIMA to five domain expert in the health domain. One of them understood his task (interaction with HIMA) after an introductory depiction, whereas the others understood their tasks after relating each step in detail. Therefore, we put all description about how HIMA works in the help option of the main menu of that system. All questions which were produced by HIMA to the domain experts are the same questions which were produced by a knowledge engineer to domain experts. This means, the answers also are same in both cases. From the answers, the knowledge engineer and the HIMA were able to extract information (construction rules) and save it in the information base. That led us to conclude the knowledge base in both cases is similar (equivalent).

The following experiments show six times when the process was carried out on the same text documents both by Text Mining Intelligent Agent and knowledge engineer. The aim of this comparison was to analyze whether the Text mining intelligent agent is capable to produce and create a correct and effective knowledge base to the users. Table 4 shows the results of the above process, the column title show us the total number of pages of text document files, the number of paragraphs for each text file,

IV. .Table Comparison between TMIA with information engineer

paragraphs	Rules didn't need modification	pages in text document	accepted paragraphs	Rules need modification	Rules extracted by knowledge engineer	Total rules extracted by proposed TMIA
7	0	2	5	5	5	5
6	1	1	5	4	4	5
8	2	3	7	4	6	7

The above table 4 results show that, the comparison of various information with respect to both knowledge engineer and propose approach (TMIA). In various paragraph contents levels (7, 6 & 8), where rule need modification in average paragraph like 4 to 6. The results from knowledge engineer having less accuracy of rules than TMIA propose approach.

V. Table Results comparison between existing vs proposed with various parameter like accuracy, precision with time extraction in seconds

Conventional method Vs Proposed method	Accuracy of Data extraction in % (2D & 3D)	Precision In %	Time of data extraction in seconds	Conventional method Vs Proposed method	Accuracy of Data extraction in % (2D & 3D)	Precision In %
Lexicon	82.83	65.40	40	Lexicon	82.83	65.40
Knowledge Extraction method	79.28	81.03	37	Knowledge Extraction method	79.28	81.03
SVM-world Extractor	84.57	84.57	35	SVM-world Extractor	84.57	84.57

In above table (Table 5) conclude that the propose method (HIMA and TMIA) obtained more accuracy and optimum precision than conventional methods like both SVM and knowledge extraction). In lexicon method having complex paragraph obtaining in various text and paragraph environment. Time of extraction is maximum in conventional methods than proposed method.

In mining system, time of execution with data extraction should complex for calculation. In our proposed method (HIMA) includes knowledge extraction, text mining extraction etc . With some intelligent agent system. In Time computation is calculated with dimentional of the data sets like 2D and 3D with some data set size. Here we evaluated with some 2D and 3D attribute data types with data sets with mining process, we evaluate the time extraction (Execution time).

VI. Table Data Table for 0.5MB Data set 2D (0.5 MB data set, 2-D)

Execution Time (seconds)	
Lexicon (Existing)	HIMA with TMIA
3	3
6	4
10	8
11	9
16	12
30	18

 ISSN: 2321-8169
32 – 43

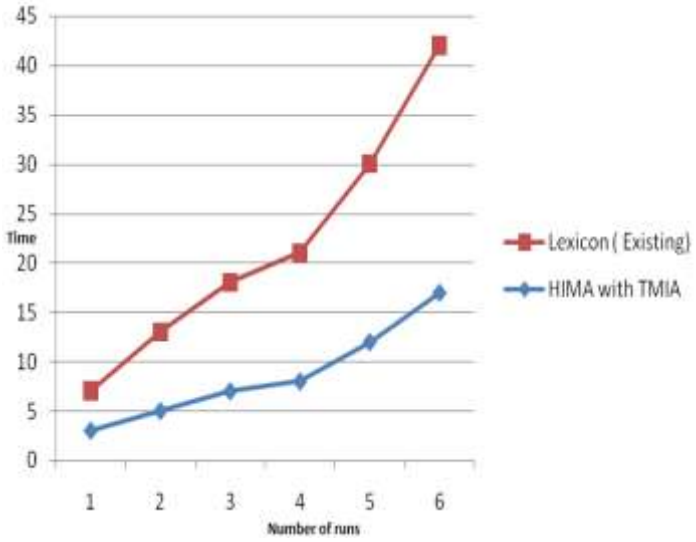

Fig. 6 Performance graph of Time execution between Lexicon with proposed approach (HIMA) Data set 2D (0.5 MB data set, 2-D)

In Fig.6 shows that performance time execution between existing (lexicons) with proposed method (HIMA with TMIA). The performance shows between numbers of runs in db with time extraction with respect to 2D data sets.

VII. Table Data Table for 0.5MB Data set 3D (0.5 MB data set, 3-D)

Execution Time (seconds)	
Lexicon (Existing)	**HIMA with TMIA**
4	3
8	5
11	7
13	8
18	12
25	17

Execution Time (seconds)	
Lexicon (Existing)	**HIMA with TMIA**
4	3
8	5
11	7
13	8
18	12
25	17

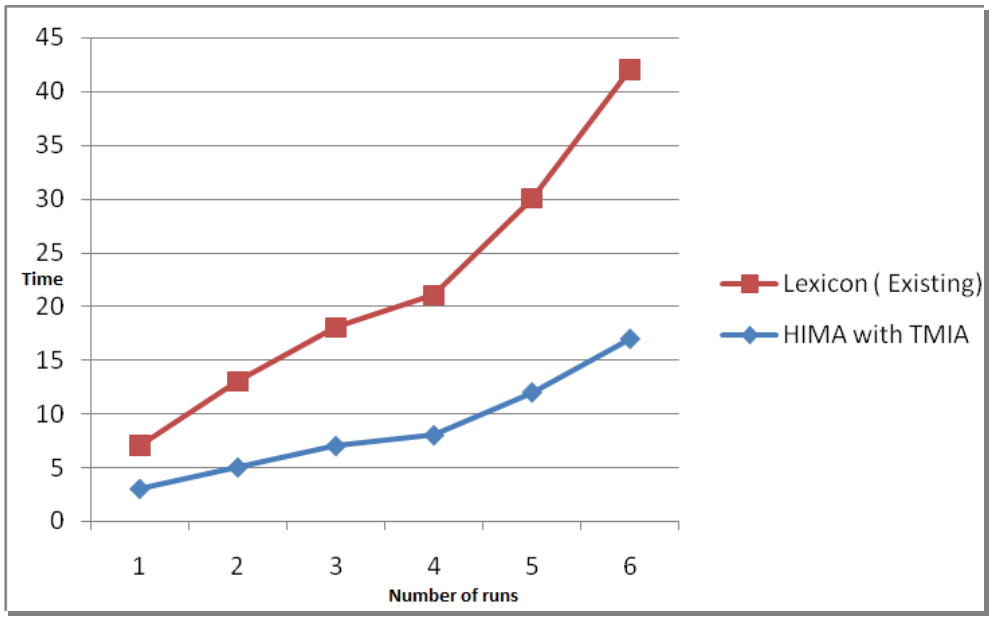

Fig. 7 Performance graph of Time execution between Lexicon with proposed approach (HIMA) Data set 3D (0.5 MB data set, 3-D)

In Figure7 shows that performance time execution between existing (lexicons) with proposed method (HIMA with TMIA). The performance shows between numbers of runs in db with time extraction with respect to 3D data sets.

Conclusion

The method presented in this work offers new approaches for acquire information from field experts and text documents based on a Hybrid intelligent Multi agent (HIMA) and using an automatic way for constructing a knowledge base in a specific diagnosis domain. In this research paper we produce expert mining as a new concept to mean extracting useful knowledge or meaningful patterns from human domain experience and we present a novel method for text mining based on a medical database, causal words, phrase structure trees, and predefined template to extract production rules from text document files. The

proposed architecture (HIMA with TMIA) can speed up construction of a knowledge base by reducing the amount of time (retrieval and execution t ime) that a domain expert may take when trying to explain his experience to knowledge engineer and reducing the time also for the knowledge engineer when he reads text documents to extract and formulate knowledge from these documents

References

[1] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration.IEEE Intelligent Systems, 18(5):16–23,2003.
[2] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), pages 39–48, Washington, DC, Aug.2003
[3] Bing Liu. Web Data Mining - Exploring Hyperlinks, Contents and Usage Data. Springer, 2007
[4] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J.Mooney, A. K. Ramani, and Y. W. Wong. Comparative experiments on learning information extractors for proteins and their interactions.Artificial Intelligence in Medicine (special issue on Summarization and Information Extraction from Medical Documents), 33(2):139–155, 2005.
[5] M. E. Califf and R. J. Mooney. Bottom-up relational learning of pattern matching rules for information extraction.Journal of Machine Learning Research, 4:177–210, 2003
[6] K.-T. Chen, C. Chen, and P.-H. Wang. Network aware loadbalancing via parallel vm migration for data centers. In Computer Communication and Networks (ICCCN), 2014 23rd International Conference on, pages 1–8. IEEE, 2014.
[7] C. Clifton and R. Cooley. Topcat: Data mining for topic identification in a text corpus. In Proc. European Conf on Principles of Data Mining and Knowledge Discovery(PKDD 99), pages 174–183, 1999
[8] Das-Neves, F., Fox, E. A. and Yu, X. Connecting Topics in Document Collections with Stepping Stones and Pathways. CIKM‟05, ACM Press, 2005, pp. 91-98
[9] M. A. Hearst and C. Plaunt. Subtopic structuring for fulllength document access. In Proc. 16th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pages 59–68, 1993.
[10] R.Hemamalini, Dr.L.Josephine Mary "An Analysis on Multi-Agent Based Distributed Data Mining System", International Journal of Scientific and Research Publications, Volume 4, Issue 6, June 2014.
[11] T. Kohonen. Self-Organizing Maps. Springer-Verlag, Berlin,1997.
[12] C. Y. Lin. Knowledge-based automatic topic identification.In Proc. Meeting of the Association for Computational Linguistics (ACL 95), pages 308–310, 1995.
[13] Lokesh Kumar, Parul Kalra Bhatia "**TEXT MINING: CONCEPTS, PROCESS AND APPLICATIONS",** Journal of Global Research in Computer Science, 4 (3), March2013,36-39
[14] http://www.scism.lsbu.ac.uk/inmandw/ir/jaberwocky.htm
[15] http://www.freepatentsonline.com/
[16] http://searchbusinessanalytics.techtarget.com/definition/text-mining
[17] Mohammed Abbas Kadhim, M. Afshar Alam , Harleen Kaur **A Multi-intelligent Agent Architecture for Knowledge Extraction: Novel Approaches for Automatic Production Rules Extraction,** International Journal of Multimedia and Ubiquitous Engineering Vol.9, No.2 (2014)
[18] J. M. Ponte and W. B. Croft. Text segmentation by topic.In Proc. European Conference on Digital Libraries (ECDL 97), pages 113–125, 1997
[19] G. Salton and A. Singhal. Automatic text theme generation and the analysis of text structure. Technical Report TR 94-1438, Dept. Computer Science, Cornell Univ., Ithaca, NY, 1994
[20] Santosh Kumar Paul, Madhup Agrawal, Shyam Rajput, Sanjeev Kumar "An Information Retrieval(IR) Techniques for text Mining on web for Unstructured data", International Journal of Advanced Research in Computer Science and Software Engineering 4(2),February-2014, pp.67-70.