International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.2 (2014), pp.95-114
http://dx.doi.org/10.14257/ijmue.2014.9.2.10

A Multi-intelligent Agent Architecture for Knowledge Extraction: Novel Approaches for Automatic Production Rules Extraction

Mohammed Abbas Kadhim[1,2], M. Afshar Alam[1] and Harleen Kaur[1]

[1]*Department of Computer Science, Hamdard University, New Delhi, India*
[2]*College of Computer Science and Mathematics, University of Al-Qadisiyah, Iraq*

moh_abbas74@yahoo.com, mailtoafshar@rediffmail.com, harleen_k1@rediffmail.com

Abstract

In this paper, multi-intelligent agent architecture has been proposed for automatic knowledge extraction from its resources (domain experts and text documents). The extracted knowledge should be stored in a knowledge base to be used later by knowledge-based systems. This article aims to produce an effective knowledge base by cooperation between expert mining and text mining techniques. Firstly, we are constructing an Expert Mining Intelligent Agent (EMIA) able to interview with domain experts for mining problem solving knowledge as production rules in a specific diagnosis domain. It is also responsible for extracting the patterns or linguistic expressions and save it in a conceptual database. Secondly, we are constructing a Text Mining Intelligent Agent (TMIA) capable of extracting production rules from a text document corpus. The achievement of that extraction can be performed by a text document categorization based on a traditional term weighting scheme (TF-IDF) and using the Stanford parser to analyze and produce a parsing tree for each sentence in that document. Then, the TMIA looks for all causal words and takes them as separation words to generate patterns and sub-patterns based on the conceptual database. Finally, the TMIA stores those patterns and sub-patterns in a pre-formatted template and displays it to a domain expert for a modification process to construct accurate production rule.

Keywords*: multi-intelligent agent, knowledge base construction, automatic knowledge acquisition, expert mining, text mining, text documents categorization*

1. Introduction

Human domain experts and natural language text documents are the main resources of knowledge for constructing knowledge-based systems. The elicitation of that knowledge of traditional methods is the bottleneck for constructing those systems. A knowledge-based system works on a knowledge base that contains the problem solving knowledge extracted from domain expert. This knowledge base is represented by some approaches of knowledge representation (production rules, frames, Bayesian networks, *etc.*) and is built by the knowledge engineer from extracted domain expert knowledge and, later, validated by an expert [1]. The extraction of knowledge directly from domain experts and text documents allows for elicitation knowledge easily and without intervention of knowledge engineers. The main focus of this research paper is the problem of extracting interesting knowledge from domain experts and text documents using expert mining and text mining techniques respectively. We are suggesting the definition for expert mining as a process of extracting useful knowledge or patterns from a human domain expert directly without the interference of a knowledge engineer, while text mining is concerned with the detection of patterns in natural language texts, just as data mining is concerned with the detection of patterns in databases [2].

Intelligent agents or software agents are programmed software entities that carry out a series of operations on behalf of a user or another program with some degree of autonomy [3].

ISSN: 1975-0080 IJMUE

A Multi-Intelligent Agent (MIA) is a collection of autonomous agents which interact with each other or with their environments to achieve one or more objectives. If intelligent agents have any type of conversation with the user, they are also called conversational agents [4, 5]. In the last decade agents have become a very popular paradigm in computing. The reasons for this popularity are their modularity, flexibility, and general applicability to a wide range of problems [6]. Recently, the use of intelligent agents has been applied in different applications such as intelligent agent user interface, autonomous agents, information retrieval, and knowledge discovery and data mining. Intelligent agents can be classified into two different categories: resident and mobile. Resident agents stay in the computer or system and perform their tasks there. Mobile agents can travel autonomously through different system architectures and platform to fulfill their jobs [7, 8].

In this paper, we propose 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 which 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 [4]. Then, this agent calls the segmentation text procedure to split the accepted text document into a set of individual sentences and using the Stanford parser to analyse each sentence to find the main concepts or patterns and relationships based on a matching process with the contents of the conceptual database. Finally, the intelligent agent stores that patterns in pre-defined template to prepare it for final formalization as antecedence-consequence rules.

The remaining structure of this paper can be organized as follows: Section 2 gives an overview of related works on using intelligent agents in information extraction and text mining techniques. Section 3 discusses knowledge engineer functions, knowledge representation approaches and how we can select problem domain in the proposed system. Section 4 presents the components of the proposed system architecture which consists of two subsections, EMIA and TMIA. The completeness and consistency of the knowledge base and the system evaluation are discussed in Section 5 and 6 respectively. Finally, Section 7 contains the conclusions of this work.

2. Related Works

In this section, we produce an overview of studies and researches related with 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.

2.1. 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 [9].

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) [4].

Ralha, & Silva, (2012) suggested a multi-agent data mining system for extracting useful information from the Brazilian federal procurement process databases used by government auditors in the process of corruption detection and prevention to identify cartel formation among applicants. Extracting useful information to enhance cartel detection is a complex problem because the large volume of data used to correlate information and the dynamic and diversified strategies companies use to hide their fraudulent operations. To solve the problem of data volume, they have used two data mining model functions: clustering and association rules as well as a multi agent approach to address the dynamic strategies of companies that are involved in cartel formation. To integrate both solutions, they have developed AGMI, an agent-mining tool that was validated using real data from the Brazilian Office of the Comptroller General, an institution of government auditing, where several measures are currently used to prevent and fight corruption. Their approach resulted in explicit knowledge discovery because AGMI presented many association rules that provided a 90% correct identification of cartel formation, according to expert assessment [5].

2.1. 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 [10].

Mooney & Bunescu, (2005) discussed methods and implemented systems for information extraction distills structured data or knowledge from unstructured text by identifying references to named entities as well as stated relationships between such entities and summarize results on mining real text corpora of biomedical abstracts, job announcements, and product descriptions. They also discussed challenge that arise when employing current information extraction technology to discover knowledge in text [11].

Delen, & Crossland, (2008) presented, discussed and evaluated the techniques used to perform text mining on collections of textual information. A case study is presented using text mining to identify clusters and trends of related research topics from three major journals in the management information systems field. They are started from the fact of text mining is automated or a semi-automated process of extracting knowledge from a large amount of unstructured data or textual data. Given that the amount of unstructured data being generated and stored is increasing rapidly, the need for automated tools to process it is also increasing. They are also proposed that this type of analysis could potentially be valuable for researchers in any field [12].

As can be seen, intelligent agent concept and text mining techniques play a main role in knowledge acquisition and information extraction as well as increasing attention for helping in knowledge elicitation. In this paper, we adopted a different cooperative approach for constructing efficient knowledge base by producing expert and text mining intelligent agents.

3. Knowledge Engineering

The process of acquiring knowledge from experts and building a knowledge base is called knowledge engineering [13]. Fox defines knowledge engineering as ‘‘the engineering discipline that involves integrating knowledge into computer systems in order to solve complex problems normally requiring a high level of human expertise’’ [14].

Knowledge engineering is one of the research domain topics which require expert's knowledge to solve. It involves knowledge acquisition, representation and optimal usage to find effective solutions to these problems. The person does this function is called knowledge engineer. Knowledge engineers and domain experts work together and use hybrid languages to solve problems and to make that knowledge more convenient for systems that require not only knowledge extraction from experts but also structure representation to be satisfied.

The knowledge possessed by domain experts is sometimes unstructured and not explicitly expressed and they have developed numerous techniques to facilitate the extraction process . The goal of knowledge engineering is to help experts articulate what they know and document the knowledge in a reusable form [13, 15]. Each stage of knowledge engineering requires skills and special expertise to produce an effective knowledge base. Although the construction of knowledge base systems is almost similar to constructing traditional systems, knowledge base systems focus on using artificial intelligence techniques for capturing and applying their knowledge. As well, those systems are characterized by high performance. This refer to their ability to implement functions with skills depending on the quantity and quality of knowledge in those systems [16, 17], the knowledge engineering process has different phases of overlapping sub-processes. The most important phase in that process is the selection of the problem domain space. Therefore, there are many criteria for selecting problem domain:

- At least an expert should exist and he/she should be ready and able to produce his/her knowledge in specific domain of expertise
- Ability to transfer skills of experts to computer programs
- The problem we want to solve can't be solved by traditional computation methods
- The problem should be of appropriate size and scope [18].

The process of knowledge engineering involves many activities which have been studied to develop the efficiency of those activities. First, knowledge acquisition includes interaction between domain experts and knowledge engineers to extract problem solving knowledge to use later. The knowledge representation activity includes selecting one or more ways for formulation or representation of acquired knowledge. Finally, the third activity is specifying the diagnosis domain of the problem solving knowledge by interaction between experts and intelligent agent which is responsible for expert mining in the proposed system.

3.1. Knowledge acquisition

Knowledge acquisition includes the acquisition of knowledge from human domain experts, text documents, sensors, or computer files. The knowledge may be specific to the problem domain or to the problem-solving procedures, it may be general knowledge, or it may be meta-knowledge (knowledge about knowledge). Failure to acquire and organize the appropriate amount of relevant knowledge reflects on the quality and quantity of the knowledge system [13, 19]. There are two types of sources of knowledge: documented knowledge and undocumented knowledge. The first type exists in different documented

sources such as books, scientific researches, databases…etc. and the second one resides in people's' minds as experience. Knowledge can be identified and collected by human sense or by machines such as sensors, pattern matchers and intelligent agents [13,20]. Knowledge exists in different organizations and it takes various forms. However, these organizations have difficulties to make systems able to acquire, retain and access it especially in specialized knowledge [15].

We can divide knowledge acquisition processes into two approaches: manual and automatic knowledge acquisition. In manual or traditional approaches, the function of the knowledge engineer is to interview or observe domain experts to elicit relevant knowledge and then code it in the knowledge base. The characteristics of this approach are slow, limited and expensive. Therefore, we aim to automating the process of knowledge acquisition as much as possible. Automatic approaches make it possible to build a knowledge base without intervention of knowledge engineers and only need experts to interact directly with that system to extract relevant knowledge from them and store it in a knowledge base. In this paper we propose two intelligent agents for that task, one for undocumented knowledge and the other for documented knowledge.

3.2. Knowledge representation

The acquired knowledge in the knowledge acquisition process should be organized and formulated using one of the approaches of knowledge representation. In fact, there are a lot of methods for knowledge representation: the most commonly used methods include production rules representation, semantic network representation, frame representation and logic form representation, and each one has its advantages and disadvantages [21]. There is another method of knowledge representation which combines two or more of the previous methods. The purpose of this combination is to produce an integrated method of knowledge representation through overcoming the disadvantages in one method by using the second one. This type of combination is called hybrid representation. For example, there are many systems which show how a frame representation can serve as a powerful foundation for a production rule representation. The frames provide a rich structural representation for describing the objects referred to in the production rules and provide a supporting layer of generic deductive capability about those objects. These frames also can be used to partition, index, and organize a system's production rules. This capability makes it easier for the domain expert to construct and understand rules, and for the system designer to control when and for what purpose particular collections of rules are used by the system [22]. The selection of a knowledge representation method is based on the nature and size of a particular problem domain, aspects of data (redundancy, noise and dependency), and the degree of confidence, as well as the completeness of knowledge in that domain [16, 23].

In the proposed system, production rules method has been used to represent the automatic acquired knowledge from domain experts and natural language text documents because this approach of knowledge representation is exactly appropriate for diagnosing problem domain which is solved by our system architecture.

3.3. Domain problem specification

The selection of a problem domain is a very important approach for constructing knowledge-based or expert systems because there are a lot of problem categories which can be solved by those systems such as interpretation, prediction, design, diagnosis… *etc.* [23].

The proposed system can solve different problems in the diagnosis domain. In other words, the proposed architecture is capable to elicit production rules in different diagnosis domain problems. Therefore, our system is a general diagnosis domain system. Now we must answer the following question: how can we determine the specific practical diagnosis domain problem?

To answer this question let us discuss the task of the expert mining agent, the function of which is to interview the domain expert to extract relevant knowledge from him directly and convert that knowledge to a set of production rules and save it in a knowledge base. At the same time, it is able to extract the linguistic expressions or patterns to save them in a conceptual database after a filtration process for each sentence entered by a domain expert to use it in text mining process. After completion of this interview, the system can determine the diagnosis problem domain. This process depends on a specialist. For example, if the expert specialist is a doctor for eye diseases, that means we will capture knowledge base which can be used for diagnosis of eye diseases.

4. Overall Proposed System Architecture

Figure 1 show the architecture of the proposed system which consists of two cooperative groups of intelligent agents. The first agent is responsible to elicit production rules by conversation with a domain expert and extract a set of linguistic patterns to construct a conceptual database. The second agent is responsible to extract production rules from natural language text documents after categorizing those document into relevant or non-relevant documents in reference to a particular diagnosis domain based on weighting term frequency (TF-IDF) for each input text document, and a text analysis process based on the Stanford parser. As well, conceptual database to find the structure of production rule. The following subsections describe in detail the components of the proposed MIA architecture.

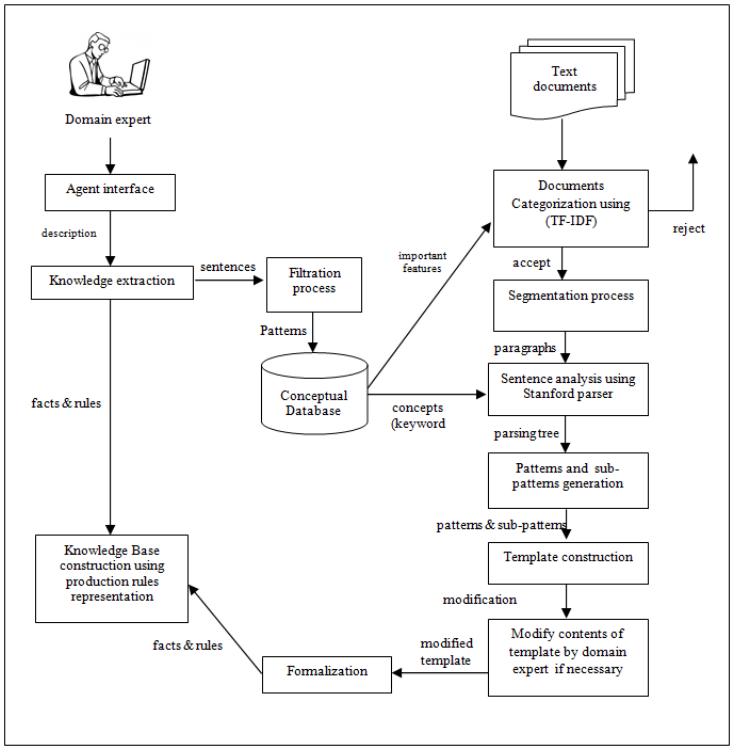

Figure 1. Proposed architecture for multi-intelligent agent system

4.1. Expert Mining Intelligent Agent (EMIA)

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 [5]. Wooldridge describes agents as computing entities which have four features: reactivity, autonomy, interaction, and initiative [24]. 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 [5, 25]. These features of an intelligent agent clearly appear in the proposed EMIA. The reactivity feature is visible when the EMIA can work in different diagnosis domain; In other words, EMIA continues to work even if human the domain experts change. EMIA 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 EMIA 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 EMIA:

4.1.1. Agent interface: One of the most important design considerations of intelligent agent systems is how we can design the agent interface as expert friendly as possible and hide the complexity of other components of the proposed agent. The intelligent agent success may be 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 EMIA to achieve the main goal of that agent.

Due to the fact that EMIA is restricted to work on the diagnosis domain using the production rules representation method, let us first discuss the diagnosis domain problem. The diagnosis is one of the general expert systems or knowledge-based systems categories which determines the cause of malfunctions in complex situations based on observable symptoms [23]; that means, in this category we have two parts: the main complex situation and a set of observable symptoms. The EMIA produces a set of questions to the expert domain and it will try to extract the main situation as well as the set of observable symptoms from the expert's answers to construct a production rule which represents a single chunk of problem solving knowledge in the knowledge base. The production rule consists of two elements: consequence or the head of the rule and the other is the antecedence or the body of the rule which should be true to satisfy the consequence part. In the diagnosis domain, the consequence represents the complex main situation and the antecedence represents the observed symptoms [26]. The EMIA requires the domain expert to enter the main situation (the disease name in the case of a clinical diagnosis domain) and also the set of observable symptoms (set of clinical observed symptoms that satisfy that disease name). The entering of the clinical observed symptoms can be done by directly entering the new symptoms or through selection from the saved symptoms in the knowledge base because of the overlapping symptoms between the new main situation (new disease) with saved situations (saved diseases) in the knowledge base. Observe Figure 2 which illustrates the flowchart of logical drawing for main steps of EMIA.

4.1.2. 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 Table 1 shows:

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.

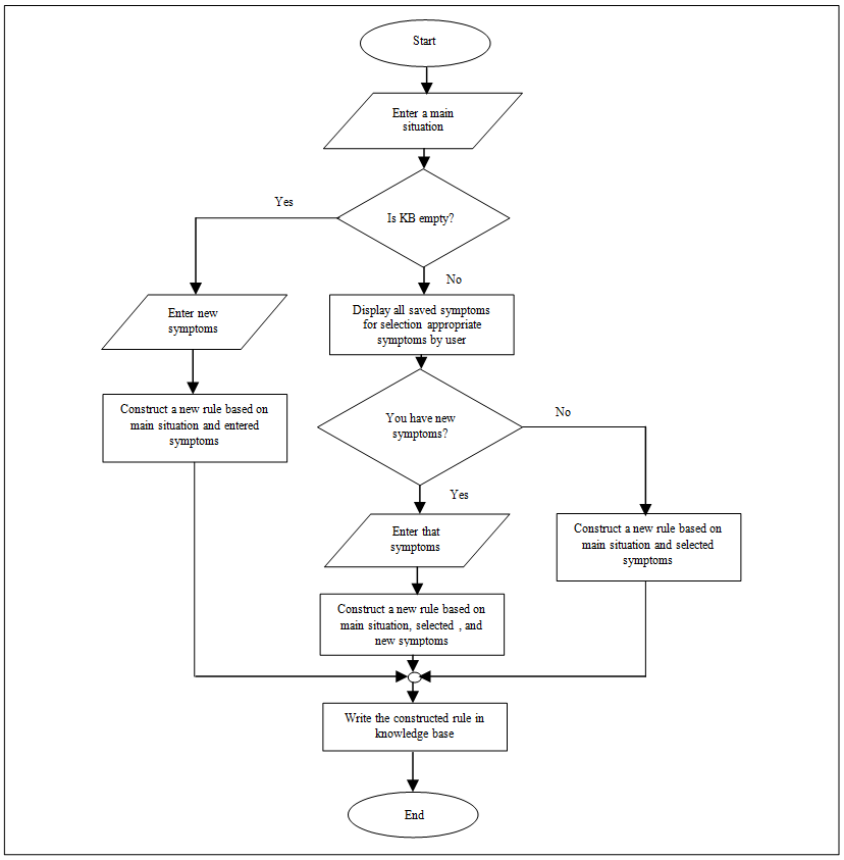

Figure 2. Flow-chart represent logical drawing for EMIA steps

Table 1. The production rule template in EMIA

Situation	Description	Pattern
Main situation	Slipped Disc	Disease(Slipped Disc)
Symptom1	low back pain in the button of the back bone	Condition1
Symptom2	leg pain in left or right one	Condition2
Symptom3	tingling, weakness, and foot senseless	Condition3
Symptoms4	little legs narcotize or numbness	Condition4
Main situation	Take drugs like Tilcotil20mg, if the situation continue do surgery operation	Treatment("slipped disc", "Take drugs like Tilcotil20mg, if the situation continue do surgery operation").

Based on descriptions and patterns in the above template, the EMIA tries to construct a production rule or a fact as clauses in Prolog, Prolog is a logic programming language which is widely used for implementation of rule-based and expert systems [27]. For the above example, the EMIA constructs the following production rule and fact in Prolog language code:

```
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")
```

4.1.3. Knowledge base construction : The knowledge base which is constructed automatically by EMIA contains up-to-date knowledge in a practical domain, because it interacts directly with the domain expert, who has up-to-date and empirical expertise in that domain. Once the production rule has been constructed in the knowledge extraction phase, the EMIA saves it in the knowledge base. The constructed knowledge base can be divided broadly into three parts:

- The condition-action knowledge base: it contains all production rules that are formulated as condition-action pairs. The action or consequence part is represented as a main situation (disease name for above example) and the condition or antecedence part is represented as a set of observable symptoms which are coded as condition1, condition2…, condition N, where N is the number of conditions.
- The textual knowledge base: it contains textual symptoms (disease textual symptoms in the case of the clinical diagnosis domain) in the form of sentences. Each sentence in the textual knowledge base represents one condition in the condition-action knowledge base.
- The treatment knowledge base: it is especially for the clinical diagnosis domain which contains the disease treatment knowledge base and it is related to the condition-action knowledge base by the disease name. That means, for each disease name in the condition-action knowledge base there is a clinical treatment in the treatment knowledge represented as a fact with two string arguments: the disease name and the treatment for that disease.

4.1.4. Conceptual database

The other function of the EMIA is to extract a set of patterns (keywords) and linguistic expressions from the expert's answers and save them in a conceptual database to use later by TMIA. In the expert's answers there are some noisy or meaningless words that do not contain knowledge. These words effect the accuracy of patterns or expressions which are extracted by the EMIA. The filtration processes eliminates all meaningless words from whole sentences in the expert's answers and stores the words and expressions which have meaning in the conceptual database as patterns or linguistic expressions. The meaningless or noisy words usually have grammar categories such as prepositions, conjunctions and interjections, and the meaning words sometimes as nouns, adjectives, adverbs and combinations between them. The

patterns and linguistic expressions which are saved in the conceptual database can be used by TMIA especially in documents categorization and patterns generation steps to help the TMIA for extracting production rules from text documents. In the expert's answers we have meaning and meaningless words. Let us suppose the meaningless words are isolation boundaries dividing each sentence into a set of meaning words as patterns (individual words) and sets of linguistic expressions (multiple words) to save it in the conceptual database. For the above example, the conceptual database contains the following patterns and linguistic expressions:

Table 2. Sample of conceptual database in proposed EMIA

Patterns	Linguistic expressions
tingling	low back pain
weakness	the button
numbness	the back bone
left	leg pain
	right one
	little legs narcotize

4.2. Text Mining Intelligent Agent (TMIA)

Text mining or text data mining is a process of deriving novel information or patterns from a collection of text documents (also known as a corpus) [12]. Zhong defined text mining as "the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge or features in text documents to help users to find what they want" [28]. Although text mining is a part of the general field of data mining, it differs from data mining. The difference between them is that in the text mining the patterns are extracted from text documents rather than from a structural database. Databases are designed to process by programs automatically while text documents are written for us to read. We do not have any programs to read and understand text documents [12].

There are many techniques that have been developed and can be used in text miming processes such as information extraction, summarization, categorization, clustering, topic tracking, concept linking, question answering, and information visualization [29, 12]. In the proposed TMIA we used two techniques: the categorization technique which is categorizing documents into relevant or non-relevant documents with reference to a particular diagnosis domain, and the other is the information or knowledge extraction technique which is used for knowledge discovery in natural language text documents. Mitra and Acharya separated the text mining process into two phases, text refinement and knowledge extraction. In the first phase, the unstructured or semi-structured text document is transformed into an intermediate form or model. In the second, the knowledge is discovered from this model by extracting patterns and mining rules [30]. Knowledge extraction is a young and rapidly growing discipline aiming to identify key phrases and relationships in text documents using a pattern matching process. The challenge in knowledge extraction is to recognize the required knowledge and extract it automatically from text documents. In this paper we proposed an intelligent agent to solve that problem. The problem is divided into set of sub-problems and the TMIA try to solve each one individually. The combination of whole solutions will produce knowledge (facts and rules) elicitation from text documents and save that knowledge in the knowledge base to use later by other systems (*e.g.*, rule-based expert systems). In Section 4.1 we discuss the four features of intelligent agents. Now let us reflect on those features in the proposed TMIA. The reactivity feature is clear when the TMIA can work in different domains of text documents after the preprocessing of those documents whether they are pertinent or non-pertinent to a particular diagnosis domain. The TMIA works autonomously to achieve its objectives and interacts or cooperates with other agents (EMIA) or human (modification of template contents by domain expert) to perform the main goal of the MIA architecture. Finally, the proposed TMIA also has initiative features while achieving

its goals. It takes initiative instead of depending only on external events. In the following sub-sections we will discuss the components of the proposed TMIA in detail:

4.2.1. Documents Categorization Using (TF-IDF): Automatic text documents categorization has been used both in the natural language process and in the organization and management of information. It is the task of assigning documents to predefined categories by computer algorithms. The text categorization approach is an important research field and is attracting increasing attention of researchers, because of the growing availability of text documents in digital format [31, 32]. The commonly used representation method for text documents is a Bag-Of-Words (BOW) where each word is treated as a feature. Features selection is defined as the process of selecting the most important features [33, 34]. In general the methods of features selection in text document categorization either depend on Term Frequency (TF) or Inverse Document Frequency (IDF). Term frequency is the number of times a particular word occurs in a document while inverse document frequency is the count of documents containing that word [33].

In proposed TMIA the problem of features selection is already solved by the conceptual database which contains the important features as patterns and linguistic expressions (phrases) which are extracted by EMIA. In text document categorization we use BOW to represent input text documents and convert each text document to a one-dimensional vector. BOW is the most widely used text representation method but it suffers from the quantity of words which is huge and it lacks to representing relationships between words [34]. To overcome these problems, we have reduced the quantity of words by eliminating all meaningless words (*e.g.*, prepositions, conjunctions and interjections) from each vector. The relationships between words are not important at this stage of the process. The proposed TMIA took the contents of the conceptual database and converted it into two sets of feature selections: one for word level and the other for phrase level. 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(t_i,d_j) refers to the frequency of term t_i in document d_j, |corpus| refers to the number of documents in the corpus, count-doc(t_i,corpus) is the number of documents in the corpus that contain the term t_i, r_i is the number of words in term t_i 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 $TF - IDF(t_i, d_j)$ is the weight of term frequency for the i^{th} term, N is the number of features (terms), S is the total number of words that compose N. Now the TMIA should make

a 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

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.

4.2.2. Sentence analysis: After determining whether the text documents corpus is related to a particular diagnosis domain, the TMIA filters it by eliminating all paragraphs which are non-relevant with that domain and processing the relevant paragraphs. In other words, the TMIA matches between the concepts (terms) in the conceptual database and each paragraph in text documents to isolate relevant and non-relevant paragraphs and processes each relevant paragraph separately by dividing it into set of sentences using a segmentation procedure. Then it isolates the sentences which have words matched with the concepts (terms) in the conceptual database as sub-patterns which will become parts of proposed observable symptoms that process in the next stages. The process will continue with other sentences in relevant paragraphs.

The entities and relations are identified by using Natural Language Processing (NLP) tools in a document [35]. The sentences are parsed for Part-Of-Speech (POS) analysis which assigns POS tags to each word in a sentence. The POS tags are used to identify the grammatical structure of sentences such as verb and noun phrase and also identify the syntactic category for each word such as nouns, adjectives,…etc. in that sentence. For POS analysis we have used the Stanford parser, which is a probabilistic natural language parser recognizing the grammatical structure of sentences [36, 35]. In the proposed TMIA the Stanford parser receives paragraph as input and convert each sentence in that paragraph into an equivalent grammatical structure tree. Table 3(a) shows us a sample of an original text paragraph relevant to a back pain diagnosis domain, and Table 3(b) illustrates the same sample but after the process has isolated the sentences matched with the contents of the conceptual database and its corresponding phrase structure tree has been generated by the Stanford parser.

Table 3(a). Original text paragraph and matched concepts with conceptual database contains

Original text paragraph
Spinal stenosis occurs when the spinal cord is compressed. This condition should be suspected in patients with ***low back pain*** that is aggravated by walking and with hyperextension of the back and that is relieved by rest or flexion of the back. These patients often have fewer symptoms walking uphill than downhill, because the volume of the spinal canal increases with back flexion and decreases with extension. Patients may also report pseudoclaudication and sciatica. Pseudoclaudication or bilateral ***leg pain*** can occur with walking or prolonged standing [37].

Table 3 (b). Text paragraph and corresponding phrase structure tree generated by Stanford parser

Text paragraph	Phrase structure tree
Spinal stenosis occurs when the spinal cord is compressed. These patients often have fewer symptoms walking uphill than downhill, because the volume of the spinal canal increases with back flexion and decreases with extension. Patients may also report pseudoclaudication and sciatica.	(ROOT(S(NP (JJ Spinal) (NNS stenosis))(VP (VBZ occurs)(SBAR (WHADVP (WRB when))(S(NP (DT the) (JJ spinal) (NN cord))(VP (VBZ is)(VP (VBN compressed))))))(. .))) (ROOT(S(NP (DT These) (NNS patients)) (ADVP (RB often))(VP (VBP have)(NP (NP (JJR fewer) (NNS symptoms))(VP (VBG walking) (ADVP (RB uphill)) (ADVP (IN than) (RB downhill)))) (, ,) (SBAR (IN because)(S(NP(NP (DT the) (NN volume))(PP (IN of) (NP (DT the) (JJ spinal) (NN canal))))(VP(VP (VBZ increases) (PP (IN with)(NP (JJ back) (NN flexion))))(CC and) (VP (VBZ decreases)(PP (IN with) (NP (NN extension)))))))) (. .))) (ROOT(S(NP (NNS Patients))(VP (MD may) (ADVP (RB also))(VP (VB report)(NP (NN pseudoclaudication) (CC and (NN sciatica)))) (. .)))

4.2.3. Patterns and sub-patterns generation: In Section 4.1.1 we discussed the relationships between the diagnosis domain and the production rules representation. Let us now discuss how we can extract the components of the production rule from text based on the result of sentences analysis stage. In this section we use patterns and sub-patterns to represent a main situation and observable symptoms respectively. The EMIA tries to identify the causal words in a process paragraph to mark them as separation words between proposed patterns and proposed sub-patterns. Typically, the causal words are used to indicate causal relationships between two objects (patterns and sub-patterns) in a paragraph. In the English language we have many causal words such as "because", "if", "report", "suffer", "symptom", "condition" and so on. Now let us suppose that all the noun phrases in the phrase structure tree before these words are proposed patterns and all the phrases after those words are proposed sub-patterns. In other words, after the complete processing of this stage the TMIA has a set of proposed patterns or main situations and a set of proposed sub-patterns or observable symptoms which produce to the next stage.

4.2.4. Template construction: In this stage, the TMIA takes all proposed patterns and sub-patterns which yielded from previous stage and all proposed sub-patterns which resulted from the sentences analysis stage. Then, it puts all of them in a predefined template which is prepared for that purpose. Table 4 illustrates the contents of predefined templates extracted from Tables 3(a-b).

Table 4. Proposed patterns and sub-patterns extracted from Table 3(a-b)

Proposed patterns (main situations)	Proposed sub-patterns (observable symptoms)
• Spinal stenosis • The spinal cord	• This condition should be suspected in patients with low back pain that is aggravated by walking and with hyperextension of the back and that is relieved by rest or flexion of the back • These patients often have fewer symptoms walking uphill than downhill • because the volume of the spinal canal increases with back flexion and decreases with extension. • Patients may also report pseudoclaudication and sciatica • Pseudoclaudication or bilateral leg pain can occur with walking or prolonged standing

The contents of Table 4 are also displayed for the domain expert to select, modify, add and delete appropriate proposed patterns and sub-patterns to capture and accept certain patterns and sub-patterns as in Figure 3 which illustrates the modification process by the domain expert on proposed patterns and sub-patterns which were produced by TMIA and sent to final predefined template as shown in Table 5 for use in the formalization process. The final predefined template consists of three components:

1. Patterns and sub-patterns: it includes a main situations and a set of symptoms in the case of a production rule description or only a main situation in the case of fact description.
2. Description: it includes all real descriptions for each situation in the first component as phrases which are modified (select, update, add, and delete) by domain expert.
3. Extracted formula: it consists of two types of formulas; main formula which represent the head of the rule (consequence) or fact, and sub-formula which represent the body of the rule (antecedence).

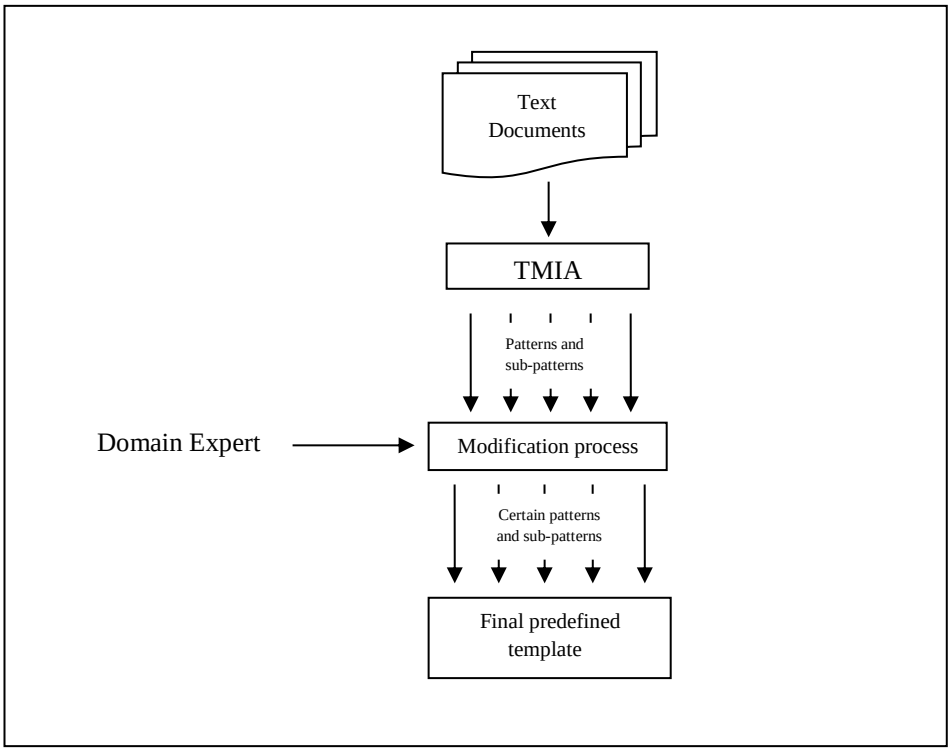

Figure 3. Modification process by domain expert on proposed patterns and sub-patterns

Table 5. Final template modified by domain expert in TMIA

Patterns and sub-patterns	Description	Extracted Formula
Main situation	spinal stenosis	disease(spinal stenosis)
Symptom5	This condition should be suspected in patients with low back pain that is aggravated by walking and with hyperextension of the back and that is relieved by rest or flexion of the back	condition5

Symptom6	These patients often have fewer symptoms walking uphill than downhill	condition6
Symptom7	Pseudoclaudication or bilateral leg pain can occur with walking or prolonged standing	condition7
Main situation	The physical therapist will help you try to reduce your pain, using stretches . If the pain does not respond to these treatments, or you lose movement or feeling, you may need surgery.	treatment (spinal stenosis, The physical therapist will help you try to reduce your pain, using stretches . If the pain does not respond to these treatments, or you lose movement or feeling, you may need surgery)

4.2.5. Production rules capture (formalization): Depending on descriptions and extracted formula from the above final pre-defined template, the TMIA tries to construct a production rule and a fact as clauses in Prolog programming language codes. For the above example, the TMIA constructs the following production rule and fact in Prolog language code:

```
disease_ name("spinal stenosis"):-
      condition5,
      condition6,
      condition7.
conditions(condition5," This condition should be suspected in patients with
                  low back pain that is aggravated by walking  and with
                  hyperextension of the back and that is relieved by rest
                  or flexion of the back ").
conditions(condition6," These patients often have fewer symptoms walking
uphill than downhill ").
conditions(condition7," Pseudoclaudication or bilateral leg pain can occur
with walking or prolonged standing ").
treatment("spinal stenosis", " The physical therapist will help you try to
                        reduce your pain, using stretches . If the pain
                        does not respond to these treatments, or you lose
                        movement or feeling, you may need surgery.").
```

Once the production rule and fact have been constructed in the formalization phase, the TMIA saves it in the knowledge base. In Section 4.1.3 we discussed three parts of the knowledge base in the proposed system: the condition-action knowledge base, the textual knowledge base, and the treatment knowledge base. The TMIA saves each part of the constructed production rule in its place in each part of the knowledge base. For example, it save the condition action pair of the production rule in the condition-action knowledge base and the textual symptoms in the textual knowledge base and so on. The new contents of the textual knowledge base part should be sent to the filtration process as sentences to extract new patterns and linguistic expressions and save them in the conceptual database. The TMIA returns to the sentences analysis stage to take new relevant paragraph and repeat all the processes again until the last relevant paragraph.

5. Knowledge Base Completeness and Consistency

Now we capture a complete knowledge base which yields from cooperation between EMIA and TMIA using the production rules knowledge representation method. When the knowledge representation in the knowledge base is completed, or is at least at a sufficiently high level of accuracy, it is ready to be used. The completeness and consistency of the knowledge base are the responsibility of the domain experts when they interact with the proposed multi-intelligent agent architecture.

The knowledge base completeness means the knowledge base is without any deficiency in its knowledge. In the proposed system we produce an effective knowledge base through an interactive EMIA with more than one domain expert to guarantee we can capture a variety of knowledge from different domain experts. The variety of knowledge comes also from the variety of resources for that knowledge (text documents). In this way we can obtain a complete knowledge base.

The knowledge base consistency means there is no conflict between the new added knowledge and the existing knowledge in the knowledge base. For example in the proposed architecture, if EMIA extracted production rule for a specific situation from the domain expert and the TMIA extracted the same situation from text documents, this situation should be deleted by the domain expert when TMIA displayed it to him in the template modification process, before reaching the formalization stage. In this way, we guarantee there is no conflict between adding new knowledge and saved knowledge in the knowledge base.

6. System Evaluation

The evaluation process of the proposed system is carried 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 performance of the proposed system we can divide the process into two parts: EMIA evaluation and TMIA evaluation.

In the traditional way, the knowledge engineer interviews domain experts to elicit problem solving knowledge and formulates it in the knowledge base. In the proposed architecture the EMIA interacts directly with domain experts to extract their knowledge and save it in the knowledge base. We displayed EMIA to five domain experts in the medical domain. One of them understood his task (interaction with EMIA) after an introductory description, whereas the others understood their tasks after describing each step in detail. Therefore, we put all description about how EMIA works in the help option of the main menu of that system. All questions which were produced by EMIA 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 EMIA were able to extract knowledge (production rules) and save it in the knowledge 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 TMIA and knowledge engineer. The aim of this comparison was to analyze whether the TMIA is capable to produce and create a correct and effective knowledge base to the users. Table 6 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,

Table 6. Experiment results comparison between proposed TMIA and Knowledge engineer

pages in text document file	paragraphs	accepted paragraphs	Rules didn't need modification	Rules need modification	Total rules extracted by proposed TMIA	Rules extracted by knowledge engineer
2	7	5	1	4	5	5
1	4	3	0	3	3	4
3	9	7	3	4	7	7
2	6	5	2	3	5	6
1	3	3	1	1	2	3
3	8	6	2	2	4	6

the number of accepted paragraphs, and so on. The most important columns in the above table are the number of rules which don't need to be modified, the total rules extracted by the proposed TMIA, and the number of rules extracted by the knowledge engineer. Figure 4 illustrates the comparison relationships between the number of extracted rules by TMIA

which didn't need modification by a domain expert and the number of rules which were extracted by the knowledge engineer from the same text documents. Let us suppose the average performance score of the knowledge engineer for extracting production rules from text documents is 100%. This means that based on the above table and Figure 4, the average performance score of TMIA for extracted production rules which didn't need modification process is 29%. This ratio is very low. Therefore, the TMIA can't depend on that rate to produce an effective knowledge base. To increase that ratio we proposed modification process.

Figure 5 illustrates the comparison relationships between the total number of extracted rules by TMIA (need and no need modification process) and the number of rules which were extracted by the knowledge engineer from the same text documents. The average performance score reached to 84%. This score of the proposed EMIA improved after the modification process by jumping from 29% to 84%. This ratio is not stable for all text documents, but it sometimes increases or decreases based on text document types, domain, and writing style in those documents. The knowledge base which was produced automatically by the proposed MIA architecture is very similar to the knowledge base which was produced manually by the knowledge engineer.

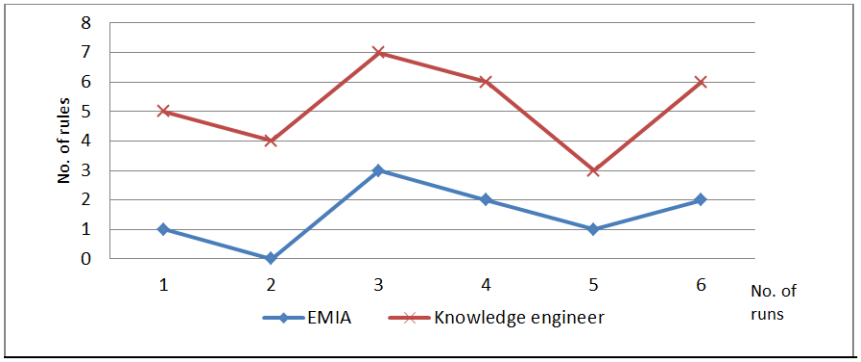

Figure 4. Comparison performance between no. of rules (didn't need modification) in EMIA and knowledge engineer

The average score may reach to around 92% between them. The knowledge base which was constructed automatically by the proposed MIA architecture can be used in expert system tools (shells) to construct expert systems in specific domains (domain of knowledge base).

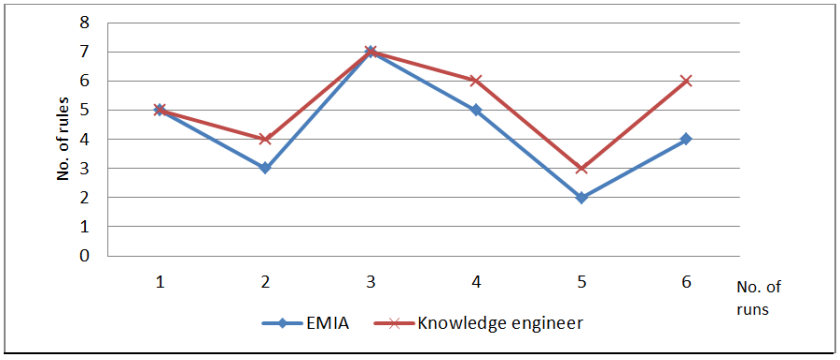

Figure 5. Comparison performance between total no. of rules in EMIA and knowledge engineer

7. Conclusions

The methodology presented in this work offers novel approaches for acquiring knowledge from domain experts and text documents based on a multi-intelligent agent 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 conceptual database, causal words, phrase structure trees, and predefined template to extract production rules from text document files. The proposed architecture can speed up construction of a knowledge base by reducing the amount of time 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.

In the system evaluation process we proved that the performance of the proposed system is optimized after the modification process by the domain expert to obtain the correct and effective knowledge base which is similar to the knowledge base that was produced by knowledge engineer. Therefore, we can use the proposed multi-intelligent agent architecture as a tool for constructing knowledge bases automatically in any diagnosis domain. That means, the proposed architecture is domain independent. The cooperation between two intelligent agents can produce a complete and reliable knowledge base because two of them can be integrated with each other. In other words, any lack in one resource of knowledge can be overcome in the other resource.

References

[1] F. Alonso, L. Martinez, A. Perez and J. Valente, "Cooperative between expert knowledge and data mining discovered knowledge: Lessons learned", Expert Systems with Applications", vol. 39, no. 8, (**2012**), pp. 7524-7535.
[2] F. Popowich, "Using text mining and natural language processing for health care claims processing", ACM SIGKDD Explorations Newsletter-Natural Language process and Text Mining, vol. 7, no. 1, (**2005**), pp. 59-66.
[3] Y. Duan, V. K. Ong, M. Xu and B. Mathews, "Supporting decision making process with "ideal" software agents – What do business executives want ?", Expert System with Applications, vol. 39, no. 5, (**2012**), pp. 5534-5547.
[4] J. Ropero, A. Gomes, A. Carrasco and C. Leon, "A Fuzzy Logic intelligent for Information Extraction: Introducing a new Fuzzy Logic-based term weighting scheme", Expert Systems with Applications, vol. 39, no. 4, (**2012**), pp. 4567-4581.
[5] C. G. Ralha and C. V. S. Silva, "A Multi-agent data mining system for cartel detection in Brazilian government procurement. Expert Systems with Applications", vol. 39, no. 14, (**2012**), pp. 11642-11656.
[6] A. Y. Seydim, "Intelligent Agents: A Data Mining Perspective", (**1999**), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.7628.
[7] D. Kim, C. Kim and K. Rim, "Modeling and Design of Intelligent Agent System", International Journal of Control, Automation, and Systems, vol. 1, no. 2, (**2003**), pp. 257-261.
[8] E. Turban, "Software (Intelligent) Agents", www.scribd.com/doc/56875420/Turban-Online-TechAppC, Technical Appendix C, (**1999**).
[9] L. Zhiping, X. Tianwei and S. Yu, "Agent-based Knowledge Management in Intelligent Tutoring Systems", The 5th International Conference on Computer Science & Education, (**2010**), pp. 578-582, IEEE, Hefei, China.
[10] R. Valencia-Garcia, J. M. Ruiz-Sanchez, P. J. Vivancos-Vicente, J. T. Fernandez-Breis and R. Martinez-Bejar, "An incremental approach for discovering medical knowledge from texts", Expert Systems with Applications, vol. 26, no. 3, **(2004)**, pp. 291-299.
[11] R. J. Mooney and R. Bunescu, "Mining Knowledge from Text Using Information Extraction", ACM SIGKDD Explorations Newsletter-Natural Language process and Text Mining, vol. 7, no. 1, (**2005**), pp. 3-10.
[12] D. Delen and M. D. Crossland, "Seeding the survey and analysis literature with text mining", Expert Systems with Applications, vol. 34, no. 3, (**2008**), pp. 1707-1720.
[13] E. Turban, J. E. Aronson, T. Liang and R. Sharda, "Decision Support and Business Intelligence Systems", Published by Dorling Kindersley (India) Pvt. Ltd., 8th Edition, (**2009**).
[14] J. Fox, "Formalizing knowledge and expertise: where have we been and where are we going?", The Knowledge Engineering Review, vol. 26, no. 1, (**2011**), pp. 5–10.

[15] S. Gebus and K. Leiviska, "Knowledge acquisition for decision support systems on an electronic assembly line", Expert System with Applications, vol. 36, no. 1, (**2009**), pp. 93-101.
[16] F. Hayes-Roth, "Rule-Based Systems", Communications of the ACM, vol. 28, no. 9, (**1985**), pp. 921-932.
[17] G. D. Bobrow, S. Mittal and J. M. Stefik, "Expert system: Perils and Promise", Communications of the ACM, vol. 29, no. 9, (**1986**), pp. 880-894.
[18] Y. P. Gupta and D. C. Chin, "Expert Systems and their Applications in production and operation management", Computers & Operations Research, vol. 16, no. 6, (**1986**), pp. 567-582.
[19] J. Yao, Y. Ma, Z. Dai, J. Niu and H. Wei, "Extracting Chemical Laws from Chinese Scientific Literature", 2009 International Conference on Artificial Intelligence and Computational Intelligence, (**2009**), pp. 287-291, IEEE computer society.
[20] W. J. Clancey, "Using the System-model-operator Metaphor for Knowledge Acquisition", IEEE Expert: Intelligent Systems and their Applications, vol. 6, no. 6, (**1991**), pp. 60-65.
[21] W. Rui and L. Duo, "The Study on Construction of Knowledge Base of Grinding Expert System Based on Data Mining", 2011 International Conference on Mechatronic Science, Electric Engineering and Computer, (**2011**), pp. 845-848, IEEE, Jilin, China.
[22] R. Fikes and T. Kehler, "The Role of Frame-Based Representation Reasoning", Communications of the ACM, vol. 28, no. 9, (**1985**), pp. 904-920.
[23] G. F. Luger, "Artificial Intelligence: Structures and Strategies for Complex Problem Solving", Published by Dorling Kindersley (India) Pvt. Ltd., 5th Edition, (**2011**).
[24] M. Wooldridge, "An Introduction to MultiAgent Systems", Published by John Wiley & Sons, LTD. England, (**2002**).
[25] L. Chen and J. Gao, "Knowledge Acquisition System Based-on Multi-agent Technology in ERP Implementation Assistant", Proceeding of the 2004 International Conference on Intelligent Mechatronics and Automation, (**2004**), pp. 152-156, IEEE, Chengdu, China.
[26] M. A. Kadhim and M. A. Alam, "To Developed Tool, an Intelligent Agent for Automatic Knowledge Acquisition In Rule-based Expert System", International Journal of Computer Applications, vol. 42, no. 9, (**2012**), pp. 46-50.
[27] N. Dunstan, "Generating domain-specific web-based expert systems", Expert Systems with Applications, vol. 35, no. 3, (**2008**), pp. 686-690.
[28] N. Zhong, Y. Li and S. Wu, "Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, (**2012**), pp. 30-44.
[29] V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1, (**2009**), pp. 60-76.
[30] S. Mitra and T. Acharya, "Data Mining: Multimedia, Soft computing, and Bioinformatics", Published by John Wiley & Sons, New Jersey, USA, (**2003**).
[31] W. Zhang, T. Yoshida and X. Tang, "A comparative study of TF_IDF, LSI and multi-words for text classification", Expert Systems with Applications, vol. 38, no. 3, (**2011**), pp. 2758–2765.
[32] C. H. Li, J. C. Yang and S. C. Park, "Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet", Expert Systems with Applications, vol. 39, no. 1, (**2012**), pp. 765–772.
[33] N. Azam and J. T. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization", Expert Systems with Applications, vol. 39, no. 5, (**2012**), pp. 4760-4768.
[34] L. Zhixing, X. Zhongyang, Z. Yufang, L. Chunyong and L. Kuan, "Fast text categorization using concise semantic analysis", Pattern Recognition Letters, vol. 32, no. 3, (**2011**), pp. 441–448.
[35] A. Coulet, N. H. Shah, Y. Garten, M. Musen and R. B. Altman, "Using text to build semantic networks for pharmacogenomics", Journal of Biomedical Informatics, vol. 43, no. 6, (**2010**), pp. 1009-1019.
[36] Jahiruddin, M. Abulaish and L. Dey, "A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora", Journal of Biomedical Informatics, vol. 43, no. 6, (**2010**), pp. 1020-1035.
[37] B. Karnath, "Clinical Signs of Low Back Pain", www.turner-white.com/pdf/hp_ may03_signs.pdf, (**2003**).

Authors

Mohammed Abbas Kadhim

He is a Ph.D. Scholar in Computer Science Dept., Hamdard University, New Delhi, India. He is also an Assistant Professor in College of Computer Science and Mathematics, University of Al-Qadisiyah, Iraq. He received his B.Sc. and M.Sc. degree in Computer Science from Babylon University, Babylon, Iraq in 1996 and 1999 respectively. He has published many research papers in Artificial Intelligence field. His research interest includes expert systems, intelligent agents, text mining, data mining, neural network, and genetic algorithm.

M. Afshar Alam

He is a Professor in Computer Science, he was Head of Computer Science Department, Faculty of Management and Information Technology, at the Hamdard University, New Delhi, India. In 1997-2000, he founded the Department of Computer Science, Hamdard University. He was also founder of Computer Centre at Hamdard University. He received his Master degree in Computer Science from the Aligarh Muslim University, Aligarh and Ph.D. from Jamia Millia Islamia University, New Delhi. His research interests include Fuzzy logic, Software engineering and Bioinformatics. He is the author of a book on Software re-engineering and over 50 publications in International/ National journals, conference and chapter in an edited book. He is a member of expert committee AICTE, DST, UGC and Ministry of Human Resource Development (MHRD), New Delhi, India.

Harleen Kaur

She gained her Ph.D. in Computer Science from Jamia Millia Islamia University, New Delhi, India on the topic of Applications of Data Mining techniques in Health care Management. She graduated from the University of Delhi, New Delhi. She has previously served as a Lecturer in Computer Science, University of Delhi. Currently, she is an Assistant Professor at the Department of Computer Science, Hamdard University. She has published numerous research articles in refereed international journals and conference proceedings and chapters in an edited book. She is a member of several international bodies. Her main research interests are in the fields of Data analysis with applications to medical databases, Medical decision making, Fuzzy logic, Information Retrieval, Bayesian networks and visualization.

