2009 International Conference on Computer Engineering and Applications
IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

Agent-based Knowledge Mining Architecture

Shaidah Jusoh[1], Hejab M. Al Fawareh [2]

[1] College of Arts and Sciences , Universiti Utara Malaysia
Sintok, Kedah, Malaysia
shaidah@uum.edu.my

[2] College of Arts and Sciences , Universiti Utara Malaysia
Sintok, Kedah, Malaysia
alfwareh@gmail.com

Abstract. - With the explosion growth of available information in texts, reading is a very time consuming task. Having an automated knowledge miner that can extract useful knowledge from the texts is very desirable. In this paper, we propose an integrated architecture for an intelligent knowledge mining processor. The architecture is based on text mining and data mining approaches, and a multi-agent system. In the text mining approach, we focus on extracting entity and concepts from the texts by using information extraction and natural language processing techniques. Of the data mining approaches, a classification and visualization techniques are applied to classify the extracted entities/concepts. The proposed architecture is designed using agent systems. Each process in the architecture is presented as an agent. Interactions between agents are conducted through messages. In this framework, FIPA ACL is proposed as an agent communication language. The main reason of using an agent technology is to decompose complex problem into sub problems and each problem is resolved independently.

Keywords: Knowledge Mining, Multivalent System, Classification, Information Extraction

1. Introduction

In the last several years, IT practitioners have agreed that there exists a continuum of data, information and knowledge. Data is mostly structured, factual, and numeric. Data consists of facts, images, or sound. When data is combined with interpretation and meaning, information emerges. Knowledge is inferential abstract that is needed to support decision making process. Knowledge can be as simple as knowing who is the president of the United States, or it can be as complex as mathematical formula relating process variables to finish product dimensions. To distinguish between information and knowledge is not always straightforward. [1] defined knowledge as “a fluid mixed of framed experience, values, contextual information, but until people use it, it isn’t knowledge”. While [2] use knowledge definition taken from [3], that the primary elements of knowledge are concepts and relationships between concepts. Basically, [4] defined concepts as ‘perceived regularities in events or objects, or records of events or objects, designated by a label’. Knowledge exists in forms such as instinct, ideas, rules, and procedures that guide actions and decision. Most researchers agree that knowledge is a human creation. Thus we can construct new knowledge by linking new concepts/entities the knowledge that we already have [5].

A large part of business’s knowledge is stored in textual documents available within the internet or intranets. The challenge is to extract that valuable information and represent its knowledge in a form that can be easily understood and reused by people or applications. Thus the field of knowledge mining has been rapidly expanding, and attracting many new researchers and users. The underlying reason for such a rapid growth is a great need for systems that can automatically deduce new knowledge from the existing information that is stored whether in structured or unstructured forms, from vast volumes of computer data being accumulated worldwide. The fields of data mining and text mining offer a promise for addressing this need. The former deal with data that is stored in a structured manner in a database system. The latter deals with data that is stored in unstructured manner within the text documents. The goal of this paper is to propose agent-based architecture for an intelligent knowledge mining process. This paper is organized as follows; Section 2 presents related areas; data mining, text mining, and multiagent systems and natural language processing. Section 3 presents our proposed architecture, and Section 4 discusses the implementation issues. The paper is summarized in Section 5.

2. Related Areas

2.1. Data Mining

Data mining is the process of automatically discovering useful information in large data repositories (databases). It is an integral part of knowledge discovery in databases. [6] quoted that "data mining is the extraction of implicit, previously unknown, and potentially useful information from data". Data mining has emerged over the past years as an effective business tool for decision support. Many successful data mining applications have been demonstrated in the area of sales analysis, fraud detection, manufacturing process, scientific analysis, etc. Many disciplines have contributed to data mining including database management systems, statistical analysis, machine learning, neural network and visualization.

2.2. Text Mining

Text Mining can be described as the process of analyzing text to extract information that is useful for particular purposes [7]. Malik [8] argued that the boundaries between data mining and text mining are fuzzy. The difference between regular data mining and text mining is that in text mining, the patterns are extracted from natural language text rather than from structured databases of facts. The use natural language processing techniques enable text mining tools to get closer to the semantics of a text source. This is important, especially when the text mining tool is expected to discover knowledge from texts. Text mining system is generally used to denote any system that analyzes large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information [9]. Clustering and categorization techniques are often used to group similar documents or queries which could serve as corporate knowledge maps. Visualization helps to reveal conceptual associations and visualize knowledge maps. Text mining is a relatively new and vibrant research area that is changing the emphasis in text-based information technologies from the level of retrieval and extraction to the level of analysis and exploration.

Text mining tools could be technologies are capable of answering sophisticated questions and performing text searches with an element of intelligence. A text mining application uses unstructured textual information and examines it in attempt to discover structure and implicit meanings hidden within the text [10]. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with. Nevertheless, texts remain the most common vehicle for the formal exchange of information. The motivation for trying to extract information from it is compelling-even if success is only partial [11].

2.3. Multiagent Systems (MAS)

MAS has been hailed as a new paradigm for conceptualizing, designing, and implementing complex software systems. Agents are sophisticated computer programs that act autonomously on behalf of their users across open and distributed environments to solve complex computing problems [12]. Sycara [13] defined MAS as a loosely coupled network of problem solvers that interact to solve problems that are beyond the individual capabilities or knowledge of each problem solver. Technically MAS provides characteristics of having advantages of computational efficiency, reliability, extensibility, maintainability, robustness, maintainability and responsiveness, flexibility, and reuse [13]. MAS is a system that contains two or more agents, at least one is an autonomous agent. The agent is a software module that has capabilities to interact with each other and perform tasks independently. Interaction between agents can be conducted through an agent communication language (ACL). MAS have been applied for various applications, including natural language processing [12] and data mining [14].

2.4. Natural Language Processing (NLP)

NLP is one of the most difficult problems in artificial intelligence area. It is the analysis on human language so that computers can understand natural language as humans do. Although this goal is still some way off, NLP can perform some types of analysis with a high degree of success. This includes parsing sentences, part-of-speech tagging, where words are classified according to their categories (noun, verb, adjective, and so on). In parsing sentences; grammar and lexicon are required. Generally NLP research work has been focused on syntactic, semantic, morphology, pragmatic and discourse analysis. When lots of valuable knowledge is hidden in texts, the role of NLP in knowledge mining has become crucial. It is a part of NLP task to identify and extract concept and entities from texts that can be used to deduce knowledge.

3. Agent-based Architecture

Knowledge miner is considered an intelligent tool whenever it can extract useful knowledge from unstructured or semi-unstructured data. Our proposed architecture is modelled by utilizing also natural language processing and information extraction techniques. The propose architecture is illustrated in Fig. 1. There are four major components in this framework; entity/concept extraction, database, mined extracted data and knowledge visualization. Fig. 2 illustrates in details of framework using MAS. In this framework, each agent is an autonomous agent. It has its own goal and performs a task independently (each agent has a specific task). Let us discuss the scenario of a multi-agent based knowledge miner system. The system consists of eleven working agents. These agents are utilized for extracting entity/concept from texts, storing and retrieving to and from database, conduct mining process on the extracted entities/concepts (data) and present knowledge to the human user. Technically entity/concept extraction focuses on identifying relevant entity/concept

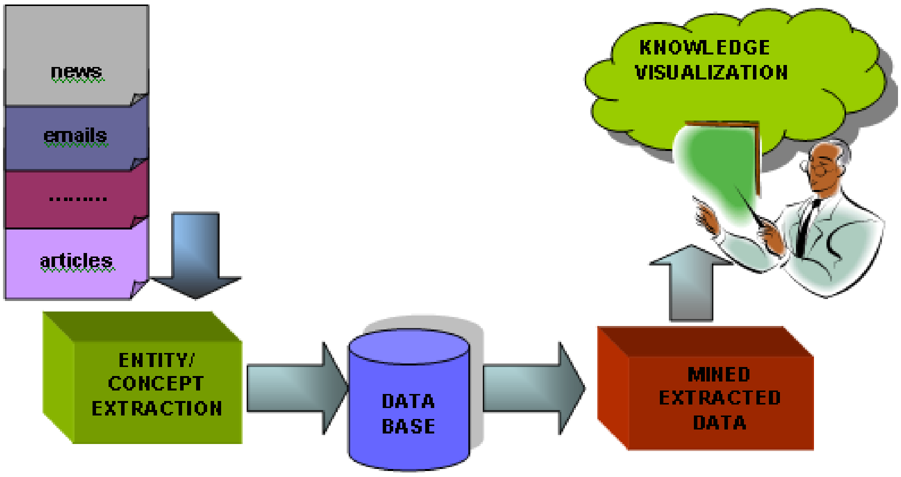

Fig. 1. A proposed framework consists of three major components: entity extraction, database and mined extracted data.

NLP is a fundamental tool in processing texts. Texts are processed sentence by sentence. Of the eleven agents, six of them are involved with texts processing. They are syntactic analyzer, lexicon, grammar, semantic, ambiguity and unambiguous agents. When a syntactic analyzer (SA) agent receives an input sentence, it conducts a parsing process to recognize structure/structures of the sentence and part-of-speech for each input words. To conduct the process, SA will communicate with grammar agent for grammar rules that are possible to be used for identifying a sentence structure. When SA agent recognizes the group of words (GW), it will communicate with the lexicon agent (LA) to identify the part-of-speech category for the recognized GW. LA is responsible to provide vocabularies to the syntactic analyzer agent. For example, SA needs to determine the word (w) “bat”. LA will check within its database whether the word “bat” is in the noun group or verb group categories. If a word can carry more than one part-speech (this is known as local ambiguity), then the LA will give all possible results to SA. SA will conduct pattern matching with the sentence structure and select the correct part-part of speech. As an example, if the word is “bat” and the word group is noun, and then the “bat” is belong to the noun part of speech category. The process can described in a procedure code below.

```
syntactic analyzer{
    if  w = noun|verb;
      then conduct pattern matching;}
```

Semantic analyzer agent (SEMA) is responsible to assign semantics to the input words. The major challenging issue in any natural language is many of natural language words are semantically ambiguous. For example the word “bat” could have a semantic of “a baseball tool“ or “an animal”. This is known as lexical ambiguity. When SEMA encounters this problem, it will pass the problem to the ambiguity agent (AMA). AMA will resolve the problem by using context knowledge and possibility theory. The techniques have been discussed in [15]. If the semantic analyzer does not encounter any ambiguity problem, it will assign the semantics and passes the assignment to unambiguous agent. Technically, it can be described in a piece of procedure below

```
semantic analyzer agent{
       if  w = ambiguous;
            then communicate with ambiguity;}
ambiguity agent {
           resolve ambiguity problem;
           communicate with unambiguous;
            else
            assign semantic;
            communicate with unambiguous;}
```

Unambiguous agent (UNA) is responsible to coordinate results from SEMA and AMA. Its task is to put the correct order of the semantic assignments. UNA is responsible to inform entity extractor agent (ETA) that the data is ready to be extracted. Based on the embedded template, ET will determine what type of entity/concept should be extracted. ETA, then will pass the extracted data to data manager agent (DMA). DMA is responsible to store and retrieve data from the database. Classification agent (CA) is responsible to mine the data in the database. Classification technique is embedded into CA for finding hidden patterns of extracted data. A common approach for classifier is to use decision trees to partition and segment records. Pattern interpreter agent (PIA), is then responsible to interpret patterns discovered by classification agent. The interpretation of the patterns is conveyed to the visualization agent. PIA is responsible to convert the interpretation into visual forms that can be easily understood by human interpreters. The visualization can make use of the domain knowledge that is locked up in people’s head. The visuals can be in the forms of line graphs, bar charts, scatter plots, icons, and so on.

4. Implementation Issues

In implementing this framework, one should have technical background in text processing, information extraction and data mining techniques. In parsing the sentence, one should have knowledge of parsing techniques; such top down and bottom up parsing techniques. One of the available algorithms, dynamic programming algorithms and has been widely. The beauty of this algorithm is it can handle a parallel top-down search efficiently. In this paper, we do not discuss in details a technique for resolving lexical ambiguity problems. It should be noted that, grammar for natural language is also ambiguous which can cause ***syntactic ambiguity***. However, as we focus on extracting entity/concepts from texts, the most related issue is local and lexical ambiguity rather than syntactic ambiguity. Although clustering and classification techniques play the same role for finding hidden patterns in data, we propose the usage of classification technique rather than clustering technique. The idea is because we have applied information extraction technique in extracting data from texts. In information extraction technique, the types of data have been known in advanced. As the framework has been constructed using a multiagent technology and agents communicate with each other using an agent communication language. The usage of FIPA ACL [18] as the agent communication language is proposed. Although there are other available agent communication languages such as KQML, FIPA ACL has been accepted as a standard language for agent communications and provides a clear semantic language.

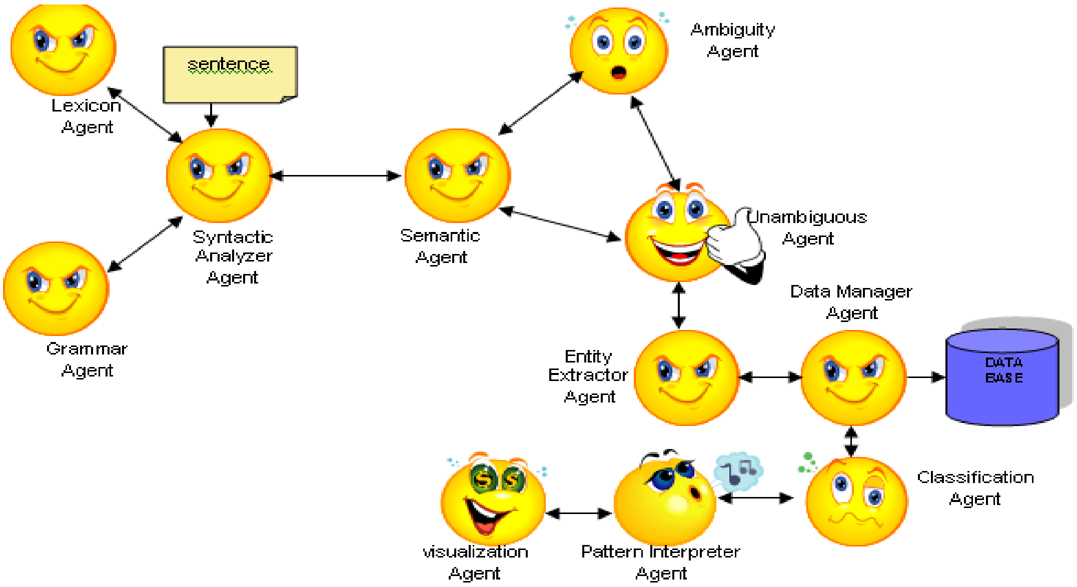

Fig. 2. Knowledge mining process is decomposed into several agents.

5. Summary

The purpose of this paper is to propose a viable framework that can be used for developing an intelligent knowledge mining system. The system is considered intelligent whenever it can extract relevance information from texts, resolve an ambiguity problem that occur in natural language words and deduce useful new knowledge. The framework is proposed to be implemented using multiagent system as it is a new paradigm of designing and developing a complex software system. The implementation issues section has given enough insight about how the implementation should be carried out.

6. Acknowledgment

This research is supported by the Fundamental Research Grant Scheme, Ministry of Higher Education, Malaysia

7. References .

[1] A. Tiwana, The knowledge management toolkit. Upper Saddle River, NJ, 2000.

[2] A.J.Cañas., R. Carff, G. Hill, M. Carvalho, M. Arguedas, T. C. Eskridge, J. Lott, R. Carvajal, Concept Maps: Intergrating Knowledge and Information Visualization in Knowledge and Information Visualization: Searching for Synergies, S.-O. Tergan, and T. Keller, Editors. Heidelberg / New York: Springer Lecture Notes in Computer Science, 2005.

[3] J.D.Novak, A theory of education. Ithaca, NY: Cornell University Press, 1977.

[4] J.D.Novak, Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations. Mahwah, NJ: Lawrence Erlbaum Associates, 1998.

[5] R.S.Michalski, Knowledge Mining: A proposed New Direction, Sanken Symposium on Data Mining and Semantic web, Osaka University of Japan, March 10-1, 2003.

[6] I.H.Witten and E.Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, 1999.

[7] A.H.Tan, Text mining: The state of the art and the challenges. In: Proceedings of Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99), 1999, pp. 65–70.

[8] R.Malik, CONAN: Text Mining in Biomedical domain. PhD thesis. Utrecht University, Austria, 2006.

[9] F.Sebastiani, Machine learning. ACM Computing Surveys **34**(1), 1–47, 2002.

[10] H.Karanikas, C. Tjortjis and B. Theodoulidis, An approach to text mining using information extraction. In: Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference, (2000).

[11] I.Witten, H., Z. Bray, M. Mahoui and B. Teahan, Text mining: A new frontier for lossless compression. In: Proceedings of the Conference on Data Compression, 1999, pp. 198–207.

[12] M.Aref, A multi-agent system for natural language understanding, International Conference on Integration of Knowledge Intensive Multi-Agent Systems (KIMAS '03), Boston, USA, 2003, pp. 36-40.

[13] K.Sycara,.P. Multiagent system, AI Magazine, Volume 19, No. 2, 1998, pp. 70-92 .

[14] W.Zhang and L. Zhang, A multiagent data warehousing (MADWH) and multiagent data mining (MADM) approach to brain modeling and neurofuzzy control. *Inf. Sci. Inf. Comput. Sci.* Vol. 167, No. 1-4, 2003, pp. 109-127.

[15] H. M. Al Fawareh and S. Jusoh, Ambiguity in Text Mining, Proceedings of the International Conference on Computer and Communication Engineering(ICCCE 2008), 2008, pp. 1172-1176.

[16] FIPA, Agent Communication Language Specification, http://www.fipa.org/repository/aclspecs.html,