CHAPTER 3

DATA EXTRACTION USING INTELLIGENT AGENT

3.1 INTRODUCTION

Analysis and extraction of important information from World Wide Web poses a phenomenal challenge to the researcher. As Information society became more data oriented with easy access of online, many new applications of data extraction came around. Research in this area is started over two decades and now reached its place as automated extraction. The foremost challenge facing the research community, in spite of more than two decades of research in the field, is designing models that achieve high accuracy of extraction.

To attract consumers, web pages are not only providing exact information which user exactly wants. The exact informations are generally covered with unwanted advertisements, pictures, videos, games, and other entertainments sources. It makes unnecessary reading and wasting time of user and the user may be tired to get exact information. This additional information deviated the user to other direction. In some cases if user wants to get exact data in same field in a continuous way then it is very difficult to find it.

Information constitutes the valuable knowledge for particular domain in web world. But due to the unstructured nature of the information, this knowledge cannot be efficiently exploited by computer machines for
automation purpose. Information Extraction is the name often used to refer to a very interdisciplinary field, which consist of using several research areas to extract data from real world data. Information Extraction is referred as the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from other structured and unstructured sources. Here are some examples

- HR firm may wish to monitor the web sites of businesses in a specific sector for available job position with salaries and locations and build and maintain a structured database containing this data to help design their pay packages.

- A market analyst may wish to monitor management changes in companies from a specified sector and get updates of the form “X replaced Y in position P of company C”.

- A researcher may wish to monitor a set of university and journal web sites for articles that claim to improve on a specific technique and to be notified with the title, authors, and a URL where the article is available online.

- An academic department may wish to monitor other universities for the promising of doctoral candidates to hire in specified areas, with related faculty being notified about significant publications by the candidates.

- A small company that assembles PCs may wish to monitor online catalogs from wholesalers of system boards, cabinets, and CPUs to detect any significant change in prices from current suppliers and bidders.
This research work is identified for the extraction of disease details in a particular region. Many diseases like Dengue fever, Bird flu, Chikungunya, and Swine flu are spreading all over India specifically in Tamilnadu. So in this search the disease details are considered as important information to the user. This work will extract the location of the disease, number of people affected and prevalence of the disease and gives the updated information of the diseases to the region wise. From that extracted data the information can be rearranged to find the disease affecting sequence and disease spreading sequence from one place to another place to prevent the disease or take necessary action for that.

The Web information extraction can be classified as artificially acquisition, semi-automatic learning style, machine learning and inductive learning methods. Depending on the source of information structure it can be categorized into the information extraction from free text, semi-structured text and structured text. Depending on extract principle it can be categorized into machine learning method, extraction method based on the natural language processing, ontology method, extraction method based on the HTML structure method and extraction method based on the XML structure method.

**Types of Data**

The IE systems can be categorized by the type of data that are used as origin:

Structured data, semi-structured data, and unstructured data.

**i) Structured Data**

This is relational data (e.g., from databases). By means of the structure of the database an exact format of the particular data is assigned.
Thus, detecting required information and the assignment of that corresponding format can be easy.

(ii) Semi-structured Data

For extracting the required information no Natural Language Understanding (NLU), like analysis of words or sentences, is required. Examples are advertisements in newspapers and job postings or highly structured HTML pages. Both XML-encoded and HTML-encoded pages contain semi-structured data, but HTML is rather more ‘human-oriented’ or ‘presentation-oriented’. It lacks the separation of data structure from layout, which XML provides.

(iii) Unstructured Data

This can be, for example, plain text. For extracting the required information the user must understand the text.

IE system will be influenced by the following factors.

Language of the text

Other than English language, some languages will require morphological and word segmentation processing.

Genre

If extracting information from speech, then the transcripts requires different techniques than text.
Text properties

The additional content in text like image and table formats need special care for handling. IR techniques may be needed for very long texts to identify the relevant sections for processing.

Task

Tasks like entity identification are relatively simple. If one wants to extract properties of entities, then the text needs to be analyzed for fragments that express the property. If the task involves extracting events, then entire clauses may have to be analyzed together.

3.1.1 Designing Approaches of Information Extraction

There are two basic approaches to the design of IE systems labeled as

- Knowledge Engineering Approach
- Automatic Training Approach

(i) Knowledge Engineering approach

The Knowledge Engineering Approach is defined by the implementation of the grammars used by tools of the IE system by a knowledge engineer. Knowledge Engineer is a person who is familiar with the IE system, and the formalism for expressing grammar for that system, who then, either on his own, or in consultation with an expert in the domain of application, writes rules for the IE system component that extract the information.
The skill of the knowledge engineer plays an important factor in the level of performance that will be achieved by the overall system. Building a high performance knowledge engineering system is usually an iterative process whereby a set of rules is written. It is also role of knowledge engineer that makes appropriate modifications to the rules, and iterates the process.

**Advantages**

- With skill and experience, good performing systems are not conceptually hard to develop.
- The best performing systems have been hand crafted.

**Disadvantages**

- Very laborious development process.
- Some changes to specifications can be hard to accommodate.
- Required expertise may not be available.

**(ii) Automatic Training Approach**

In Automatic Training Approach, it is not necessary to have an expert on hand with detailed knowledge of how the IE system works, or how to write rules for it. It is necessary only to have someone who knows enough about the domain and the task to take a corpus of texts, and annotate the texts appropriately for the information being extracted. Typically, the annotations would focus on one particular aspect of the system’s processing.

Once a suitable training set has been annotated, a training algorithm is run, resulting in information that a system can employ in analyzing novel texts. Another approach to obtaining training data is to interact with the user
during the processing of a text. The user is allowed to indicate whether the system’s hypotheses about the text are correct, and if not, the system modifies its own rules to accommodate the new information.

**Advantages**

- Domain portability is relatively straightforward.
- System expertise is not required for customization.
- Data driven rule acquisition ensures full coverage of examples.

**Disadvantages**

- Training data may not exist and may be very expensive to acquire.
- Large volume of training data may be required.
- Changes to specification may require re-annotation of large quantities of training data.

**Considerations for an Approach**

Different number of factors that influence the decision to utilize a particular approach for a particular domain of application.

- The availability of training data
- The availability of linguistic resources
- The availability of knowledge engineers
- The stability of the final requirements
The level of performance required

(i) The availability of training data

If the required training data is available or cheaply and easily obtainable, the automatic training approach can be considered. For complex domain level tasks, where the extraction task is much slower, more complex, more expensive, or requires exclusive domain expertise, the KE approach may be favored.

(ii) The availability of linguistic resources

If linguistic resources like lexicons and dictionaries are available developing rules by a knowledge engineer may be possible. Otherwise, it may be necessary to rely on training from the annotated corpus.

(iii) The availability of knowledge engineers

If there is no experienced knowledge engineer, the automatic training approach should be chosen.

(iv) The stability of the final requirements

If the user requirements change, it is often easier to make minor changes to a set of rules than to re-annotate and retrain the system. However, other changes in requirements may be easier to accomplish with a automatic training system.

(v) The level of performance required

Human skills impact for a lot. The best performing systems for various IE tasks have been hand crafted. The performance of automatically
trained systems depends on the quantity of available training data. If the data is enough for the system, the automatically training approach can achieve equivalent results to the Knowledge Engineering approach.

This research work is followed by automatically training approach. It provides a scope to adopt agent methodology easily.

3.2 TECHNICAL ISSUES IN INFORMATION EXTRACTION

3.2.1 Necessity of Information Extraction

Search engine helps the people for retrieval and use information, search and connect with others in the World Wide Web. WWW is not completed and successful without search engine. But search engine has some problems like

- High matching ratios with low precision.
- Under matching or no match.
- The retrieval results are highly sensitive to the words.
- The retrieval results are single page.

Information Extraction express the Internet content, and develop technology that is based on computational linguistics to solve these problems. But the task still seems to be too difficult. So there is a challenge and opportunity of information Extraction in each field. Challenge and opportunity of Information Extraction is concerned in two cases.

1) Knowledge management

2) Question answering system
The purpose of IE is to perform more advanced knowledge management that can realize as follows:

- Use of concept space to organize knowledge according to the meaning of the text.
- To perform knowledge maintenance with automatic programming code that has the function of compatibility check and new information extraction.
- Question answering system substitute for key words matching. i.e. To retrieve, extract and express knowledge with human support.
- Specify the user’s limitation of browse certain part of information even part of the document.

Question Answering (QA) is the task of choosing the brief and exact answer from a finite text set. QA is composed of three stages of processing:

- The problem analyzing stage
- The relevant documents searching stage
- The answer extracting and information sorting stage.

Problem analysis is to analyze the question and judge its type, pick out or deduce key words for searching of answer. Relevant documents searching are to search through source information for related documents with the key words which get from previous stage. Answer extraction and information sorting is to extract candidate answer from the related documents and sort them according to the defined standards.
With the fast shine of the Internet, it became a tremendous information service web which is globally-sited and consists of various data resources. Nevertheless, the additional information on the web page has set the IE into great difficulty. Because of the non-structural and disorder characteristic of information, user usually set an approach of full-text research to find required information, which makes the concerned webpage abound with a large quantity of unwanted additional information. Thus, to mix the useful information with useless one together makes it more difficult to pinpoint the exact information.

Therefore, in order to search the required information faster and more precisely, it is necessary to search, extract, analyze and arrange a sea of Internet information data, during which the key procedure is the information extraction. In such a way, it is convenient for the user to find what they really want among the massive information.

The objective of information extraction is that one can distinguish exact information from semi or non-structural webpage and extract it into reasonably structured and distinct form. Thus, how to extract required information from the webpage has become an important research in the field of Internet information.

### 3.2.2 Challenges in Information Extraction

The information extraction is faced with a following challenge for extracting the data from the web page. It is mainly happening in the following aspects.

- In general, information extraction is highly domain dependent. A particular system is built to solve a specific domain. However, the overall aim in the research world in the
development of information extraction is to develop a system which can easily extract in common or switch from one domain to another and can be applied to different extraction tasks without much effort.

- IE deals with identifying not only named entities but also identifying relationships between those entities and events as well.
- Not only natural language text is considered as unstructured data. Video and image can be identified in that way as well.
- The script of web page makes the web resources not visible.
- All the interactive updates occur at the same page in the website.
- The traditional information extraction is through page source of links and labels.
- The page of information display and web source of information is not consistent.

3.3 INFORMATION GATHERING

Millions of heterogeneous websites offering various services, which makes the Internet unintelligible for any single user conducting a required task. Therefore, serviceability of the Internet depends to a large degree on the amount of automated uniform user friendly services, systems, and tools. So for any specific task for the user on Internet, the user has to do a preprocessing task to get the exact result for the query. A traditional information extraction system should be capable of gathering information
from arbitrary heterogeneous websites and offer intelligent information services on its own based on information gathered.

The other side of the Internet growth is evident when the user wants to access relevant information. One of the first solutions to cope with information overload was a search engine. Typically, a user provides a couple of keywords, and a search engine finds links to relevant sites. A user must still follow the links manually, and then browse for the required information some of the links direct to vast databases. This forces the user to get acquainted with new interfaces, and slows down the user even more.

In addition, this methodology favors largest information providers due to name-branding and time optimization. Finally, this largely reduces the basic purpose of the Internet. To further complicate matters, a user is unaware of changes in already visited sites unless he or she revisits the sites frequently. So before searching exact information in domain dependent information extraction, a preprocessing such as information gathering has to made that extends the existing approaches by providing exact information to the user.

The information extraction system extracts domain specific information from natural language text. The domain and types of information to be extracted must be defined in advance. The system often focuses on object identification, such as references to people, places, companies and physical objects. It also closely involved in finding the relationship between the extracted information based on evidence in text. It is traditionally applied in situation what it is known in advance which kind of information to be extracted from a text. Information Extraction does not present the user with
entire documents, but it extracts textual unit/elements from the documents, called text regions.

The sample block diagram of the proposed information extraction technique is shown in the Figure 3.1. It has the logical link with search engine, Information gathering system and extraction system. Any commercial search engine can be connected with this system. In this work, the Google search engine is connected with the information gathering system and the important information is extracted based on the following steps.

1) The keyword is given to the search engine.

2) The search engine finds the related links. It finds the documents $d_i$ from the link and sends it to the Information Gathering System.

3) The Information Gathering System then performs the tasks such as named entity recognition and extracts the required data which the user need to process and collects the required information in a database.

4) The Data Transfer Agent transfers the required data collected by the Information Gathering System to the Extraction System.

5) The Extraction System would then perform time interval splitting, locative disease pattern, disease affecting sequence and disease spreading sequence.
In this model Information gathering system will be connected with Internet and getting related link form the Internet through search engine. The extraction system is connected with information gathering system to get the related information form information gathering system. Data transfer agent is designed to transfer the related information to extraction system. JADE agent framework is interfaced with this system to extract the information and getting the required information from the different web pages.

The architecture of the proposed Information Extraction system is shown in Figure 3.2. The data given to the information gathering system depends according to the search engine types. Preferably Google search engine is being used with this system. The related information from different web links is extracted and transferred to the information gathering system. To store the finalized data the designed database is being used. The intermediate answer for the requested information is getting iterated with the information sequence.
The final information is extracted following mining algorithm which includes Named entity recognition, extraction of location and the persons affected by diseases.

3.3.1 Information Gathering System

Information gathering is widespread provision of distributed, structured and semi-structured information resources such as the World Wide Web. In order to obtain important information that meets the user requirements, the information can only be collected from a number of different sites. The domain dependent information gathering system of this
research performs the operations such as named entity recognition, gathers the names of the locations, gathers the number of persons affected, and the time period of a particular disease affected in a particular location shown in Figure 3.3.

![Figure 3.3 Information Gathering System](image)

**Figure 3.3 Information Gathering System**

This work is mainly focused on the exact details of specific diseases which are in the form of text. The text documents $d_i$ are the documents which contain the information in the form of text. These text documents are used to extract the required data for our proposed technique. The text documents are obtained from the search engine by giving a keyword. The search engine would give a list of links based on the keyword. The text documents $d_i$ are collected from those links separately to obtain our required
data for processing our proposed technique. The sample text document with its contents as follows:

\[ d_i = \{c_1, c_2, c_3, \ldots, c_n\} \]  \hspace{1cm} (3.1)

Here,

\[ d_i \] - Number of document. Where \( i = 1, 2, 3, \ldots, m \)

\[ c_j \] - Contents in the document. Where \( j = 1, 2, 3, \ldots, n \)

### 3.3.2 Named Entity Recognition

Documents are containing text or not, mostly they are formed by a kind of structure of metadata. In semi structure or unstructured case, the equivalent label of the text has to be found. In both these cases, it does not require a huge amount of effort to extract the information. For instance, it might be sufficient to know a value separator for each field. The other names of Named Entity Recognition are Entity Identification and Entity Extraction.

The process of analyzing the text makes the named recognition difficult that depends on the type of text. In English, upper and lower case makes relatively easy to recognize the exact match that a sequence of words is an actual text. The key problem is identifying what kind of a name it is. If the text is all one case, as may well be the situation when one is processing the output, simply recognizing that a word is a name may be difficult given the considerable overlap between proper names and ordinary English words. Many approaches have been identified for named entity recognition which are listed below.

- Dictionary-based technique
- Repeated segments technique
• Boundary technique
• Pattern technique

(i) Dictionary-Based Technique

The designer has to store the related groups of text he wishes to identify in a file called a dictionary. The pattern matching is done to identify where in which document an entry in the dictionary appears. The dictionary should have a flat, list-type structure of text.

(ii) Repeated Segments Technique

All the repeated texts are identified with certain heuristics at the beginning of the data to avoid their starting with functional words without external resources.

(iii) Boundary Technique

A preliminary phase, based on a syntactic lexical-category form of dictionary, intended to label each form in a text with a grammatical label. The next phase runs linearly from start to end, through the whole text to conserve the sequences of lexical forms which are not associated with authorized boundary.

(iv) Pattern Technique

A preliminary phase is identical to that of the boundary technique. A subsequent phase runs through the text from beginning to end and identifies sequences of forms which correspond to series of sequences of specific categories such as Noun + Noun, Noun + of + Noun or Adjective + Noun.
This research work is implemented by dictionary based technique of named entity recognition. Here it categorizes the names of persons, affecting sequence, locations of disease. The other names of Named Entity Recognition are Entity Identification and Entity Extraction. It has been considered as three attributes to process the proposed technique.

### 3.3.3 Extraction of Location

The location is extracted from the document $d$ based on the following procedure:

**Step 1:** A database $ld$ has to be created which contains the name of places.

**Step 2:** The location database $ld$ is then compared with the contents $c_j$ of the document $d$ to identify the name of the place in that document $d$.

**Step 3:** Similarly, the location database is compared with all the documents $d_i$ separately to find the place mentioned in those documents.

A sample dataset $ld$ that contains the name of the places and the process of identifying a location $L$ from a document is as follows:

$$ld = \{c_1, c_2, \ldots, c_n\} \quad (3.2)$$

$$L = \begin{cases} d_i c_j, & \text{for } d_i c_j = ldc_m \\ nil, & \text{else} \end{cases} \quad (3.3)$$

where,

$ld$ - Database with the names of the location

$L$ - Extraction of location from a document
3.3.4 Extraction of Number of Persons Affected

The procedure of extraction of number of persons affected from the documents $d_i$ is as follows:

**Step 1:** A database $pd$ has to be created which contains the word ‘affected’ and the related words which gives the meaning of the word ‘affected’ or ‘diagnosed’.

**Step 2:** Check this database $pd$ with the document $d_i$ and if the document $d_i$ contains the word which is in the database $pd$.

**Step 3:** Check the preceding word $pdc_{k-1}$ and the following word $pdc_{k+1}$ of that word which is in the database and in the document.

a) If the preceding $pdc_{k-1}$ or following word $pdc_{k+1}$ is in numerical,

b) Take it as the number of the affected persons.

**Step 4:** Repeat from step 1.

A sample dataset that contains the word ‘affected’ and the related words which gives the meaning of the word ‘affected’ or ‘diagnosed’ is as follows:

$$pd = \{c_1, c_2, \ldots, c_n\}$$  \hspace{1cm} (3.4)

where,

$pd$ - Database contains the word ‘affected’ and the related words
The process of identifying the number of persons affected is as follows:

\[
P = \begin{cases} 
  pdc_{k-1}, & \text{if } pdc_{k-1} = \text{numerical and } d_i c_j = pdc_k \\
  pdc_{k+1}, & \text{if } pdc_{k+1} = \text{numerical and } d_i c_j = pdc_k \\
  nil, & \text{else}
\end{cases}
\]  

(3.5)

where,

- \( P \) - Extraction of number of persons affected
- \( pdc_{k-1} \) - \( k \)-1\textsuperscript{th} content in the database \( pd \)

### 3.3.5 Prevalence of the Disease

The prevalence of the disease is the time period which the disease exist in a particular area. It can be identified by the date mentioned in the document of the affected area.

Each set of extracted information is added, for instance, as a record to a table of a designed database in order to maintain the data that can be applied to the result dataset later. So the order of derived data has to be stored in a more structured database which is easily accessible for applying different analyzing techniques.

### 3.4 GATHERING DATA

The important information of disease like identifying the location, the number of people affected and the time period of the prevalence of disease will be extracted from web and stored in a designed database. This database is then used for the further process. In further, the data will be rearranged as per the user requirement. The Table 3.1 shows the sample database which contains the gathered information.
Table 3.1 Sample database with the gathered information

<table>
<thead>
<tr>
<th>Disease</th>
<th>Location</th>
<th>No. of Persons Affected</th>
<th>Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Madurai</td>
<td>80</td>
<td>Mar 2011</td>
</tr>
<tr>
<td>D2</td>
<td>Tirunelveli</td>
<td>71</td>
<td>Jul 2011</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dn</td>
<td>Salem</td>
<td>92</td>
<td>Nov 2011</td>
</tr>
</tbody>
</table>

3.5 INTELLIGENT AGENT FOR EXTRACTION

The agent based information extraction is attracting ever-increasing attention from both agent and data mining communities. The interaction and integration between agent and extraction can greatly complement and strengthen each side of both communities. Some complicated challenges in either community may be effectively and efficiently tackled through agent based methodologies.

The approach to the contents of an extensive set of accumulated information is very important concern nowadays. User’s expectations on data become increased when the data is not in the form of direct text. In many cases the result is in need of some help from an expert or other system on the topic. Consequently, there is the need for an intelligent agent to interpret and help the complex information the user provides and produce the concrete answers related to the existing contents of the set of domain.

Intelligent Agents, also known as Software Agents, or Multi-Agent Systems Turban & Aronson (2001), are programmed software entities and
embedded with programming tool that carry out a series of operations automatically on behalf of a user or another program. It has some degree of independence or autonomy, using some knowledge or representation of desires of the user. If an intelligent agent keeps any kind of conversation with the user, they are also known as Conversational Agents and data transfer agent which transfer data from one system to another.

The main problem of most intelligent agent is lack of flexibility. Agent can react well to correct questions, but the solutions are far of being too satisfactory when problems are vague or imprecise. In addition and also related to this lack of flexibility, many of these agents do not generate more than one result. It is essential that a user has the possibility of choosing among different options.

Information extraction agent identifies and selects the named entities relevant to the specific task of the relationships between them and events in which domain they have to be participated in the natural language text in order to make it more accessible for further manipulations.

Information is a collection of many terms and important words. In those information, a meaningful or worth terms have to be extracted and the text indexed. It is important for efficient information extraction that indexing is defined very well by appropriate terms about all data. But it is not easy to extract proper text about a document, it may be in proper for manual construction of indexing even though enormous time and labor is required for such indexing.

**General design issues**

The following factors must be considered for agent in recognizing matching patterns.
• Any textual content is characterized by basic language features like orthography, morphology, position in a sentence.

• In addition, content might be recognized by its constituents. For example, for money identification a currency can be a key feature.

• Sometimes the textual content can be clearly identified by elements which are not part of the extracted data but which appear straight before or after it within content. For instance, title abbreviations such as Mr or Dr, which occur prior to one, two or three capitalized words, give the impression that these words are part of a person’s name. This is a classic example of person’s name identification.

• Besides, a content to be extracted might depend on information which has been identified before.

3.5.1 Multi Agent Systems

Multi Agent Systems (MAS) development framework mainly focuses at facilitating the development and administration of agent-based Technologies. Currently supporting tools, such as JADE, offer huge possibilities but they are generally linked to a specific technology. JADE is specifically interfaced with java.

Currently, multi-agent systems are being applied as a solution to a wide range of problems, such as planning, scheduling systems, real-time control, robotics, and more industrial domains. This extension in the use of MAS is also captured in software engineering models based on the use of autonomous agents to solve complex and distributed problems, the definition
of methodologies, the use of programming languages or even the adoption of standards

In the current trend, JADE is the most widespread agent-oriented middleware. JADE can be defined as a distributed and modular framework that facilitates the development of agent-based applications. This agent framework implements the life cycle and the management logic of the agent by providing administrative tools to deploy, monitor, and debug multiagent systems. Researchers commonly use some existing framework in order to do the final deployment and not to spend too much time in management and communication issues.

The main advantages of this model are an easy configuration and the administrative facilities derived from the agent middleware. Nevertheless, the main drawback is that the developer will be conditioned by the characteristics of the chosen framework. Most current agent-oriented middleware although developed with the idea of providing general purpose tools applicable to a wide range of applications, have several limitations that are directly related to their particular developments. Due to hardware or resource consumption constraints most of agent based tools are linked to a technology that limits the expansion to some platforms. Within this context, several application domains, such as the systems that need short operational time requires solutions with high performance that overcome the previous limitations.

In this context, the FIPA (Foundation for Intelligent Physical Agents) committee defines the most widespread proposal within the multi agent field. The Foundation for Intelligent Physical Agents is an IEEE committee aimed at promoting agent-based technology and interoperability between agent-based applications. Due to this interoperability, it is mainly utilizing in distributed environment. FIPA specifications provide a set of
standards for agents to interoperate at different levels. One of the more relevant documents is the FIPA Abstract Architecture Specification. This document and its derived specifications define the abstract architecture proposed by FIPA for the development of MAS which is shown in figure 4.1. The main advantage of importing the set of FIPA standards is providing specifications for heterogeneous and interacting agents and agent-based models.

![Figure 4.1 FIPA agent abstract architecture](image)

The FIPA Agent Management Specification establishes the agent management model of the multi agent platform, including the basic FIPA management services, the management ontology, and the message transport model. The first relevant service is the Directory Facilitator (DF), which is responsible for providing agents with a yellow pages service. Next, the Agent Management System (AMS) act as the platform manager and maintains a directory with the authorized agent identifiers (AIDs) of the agents registered with the platform. Besides providing the agents with a white pages service, the AMS also manages the agent life cycle. Finally, the Message Transport
Service (MTS) provides support to communication between agents through Agent Communication Language (ACL) messages.

### 3.5.2 Java Agent DEvelopment Environment

Java Agent DEvelopment Environment-JADE is a software platform that provides basic middleware-layer functionalities which are independent of the specific application and which simplify the realization of distributed applications that exploit the software agent abstraction. It supports development of interoperable intelligent multi-agent systems and that is distributed under an Open Source License. A significant value of JADE is that it implements this abstraction over a well-known object-oriented language. Java provides a simple and friendly API for JADE. JADE is used by a heterogeneous community of users both in research purpose and in industrial domains.

The agent environment can evolve dynamically with peers, i.e. agents, which can appear and disappear in the model or system according to application needs and requirements. The peer Communication, regardless of whether they are running in the wireless or wired network, is completely symmetric, each peer being able to play both the initiator and the responder role.

The JADE framework organizes applications as structured ensembles of software components models, which belong to two categories:

**Agents**: These are the peers mentioned above, exhibiting autonomy and communicating through asynchronous message passing.
**Services**: These are non-autonomous components, which can execute on a single system or cooperatively on multiple systems, and whose operations can be triggered by agents.

The JADE contains one container by default and it is named as main container. The main container has three agents by default. They are AMS (Agent Management System), DF (Directory Facilitator) and RMA. If we need more agents, we can create ourselves within that container and also if we need more containers, we can create it in the JADE. The agent management system influences supervisory control over access to and use of the agent platform. In a single platform only one AMS would exist. The agent management system gives white page and life cycle service, maintaining a directory of agent identifiers (AID) and agent state. To get a valid AID each agent should register with an AMS. The default yellow page service in the platform is provided by the directory facilitator agent.

Based on design choices, JADE was implemented to provide developers with the following ready-to-use and easy-to-customize core functionalities.

- A fully distributed system inhabited by agents each running as a separate thread, potentially on different remote machines, and capable of transparently communicating with one another.

- Full compliance with the FIPA specifications.

- Efficient transport of asynchronous messages via a location-transparent API.
• A simple and effective agent life-cycle management.
  
• Support for agent mobility.
  
• A subscription mechanism for readily available agents, and even external applications.
  
• A set of graphical user interfaces to support programmers when debugging and monitoring.
  
• Support for ontologies and content languages.
  
• A library of interaction protocols which model typical patterns of communication oriented toward the achievement of one or more goal
  
• Integration with various Web-based technologies including JSP, servlets, applets and Web service technology.
  
• Support for J2ME platform and the wireless environment.
  
• An in-process interfaces for launching or controlling a platform and its distributed components from an external application.

3.5.3 Information Extraction Agent

To extract data, a frequency of term is computed and key word distribution of each term selected by using stemming, synonym for extract meaningful texts in a document. Using the extracted information with frequency and distribution, a profile with the index is constructed. And then the searching can be done for many documents using this profile for
information extraction. The criterion of term’s location can be defined by various type, document line or sentence.

Extraction system on web may utilize multi-agents to handle communication and messaging as part of the searching in different web page. Multi-agents are intended for communication and cooperation in which they have the ability to behave socially, to interact and communicate with other agents like exchange information, receive instructions and give responses and cooperate when it helps them fulfill their own goals. The role of agents for information management in distributed environment, which include resource discovery, information integrity and navigation assistance is perceived to be important. For efficient information extraction, it is important that keywords are defined very well as appropriate terms about all information in the Internet or distributed computing systems.

The agent receives a user’s information extraction request and extracts the text with frequency and distribution using the keywords of the user, and then the agent constructs profile of the information with the keywords, key paragraph, and place of the document. And then the search will be easy for many documents using the profile for information extraction and browse the document.

The data transfer agent is used to transfer the data from one system to another. The collected information which is stored in the database is transferred to the mining system through the data transfer agent. The data which we transfer through the data transfer agent will reach its destination system securely and without any loss.
3.6 CONCLUSION

Information extraction system through search engine provides exact information to the user. For automated and autonomous activity, intelligent agent is designed with the extraction system by JADE framework. The extraction can be done in two phases. First gathering data from the search engine and next extracting information from the data gathering system. The extracted information is stored in the data base for the further analysis of the information and the efficient information extraction.