CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION

In this chapter, the related work for this research work and existing techniques for information extraction by using intelligent agent are discussed. It is observed that the intelligent information extraction is related and involved in intelligent agent, information retrieval, web mining and the text mining where the data is preprocessed and extracted by the system.

2.2 AGENT BASED TECHNIQUES

In the recent technology era many innovative technologies are moving towards efficient and intelligence that are using intelligent agent framed tools for many real world applications. Many research works are encouraged and developed the agent frameworks to achieve this features. Rather than expecting developers to develop the core infrastructure by themselves, it is convenient to build agent based systems on top of an agent-oriented middleware that provides the domain-independent infrastructure.

Steven Haynes et al (2009) designed intelligent agents capable of explaining themselves which includes ontological, mechanistic, and operational explanations. These designs can help to develop better agents which support creating more usable and more affordable intelligent agents by encapsulating prior knowledge about how to generate explanations in concise
representations. This system makes use of a knowledge base and algorithms to carry out their responsibilities. A knowledge base includes domain and problem-solving knowledge specific to the agent’s tasks. This design captures, reifies, and communicates the accumulated knowledge of intelligent system developers and researchers for explaining the structure and behavior of agents. Intelligent agents that can explain themselves expose the knowledge feed into their structure and behaviors and make them more efficient resources for their users and learners.

Nicholas Jennings et al (1998) presented an overview of research and development activity of agent based technology and the Multi-Agent System (MAS). The work identifies various threads of work that exhibits research and development activity in agent based systems. Initially the agent is defined and directed by the field of Artificial Intelligence. In this field the symbolic representation of agent’s environment is represented by first order predicate logic. Then the agents are defined by using object oriented methodology. By considering this scenario, an agent has more advantages over objects. It differs with respect to the notion of flexible autonomous behavior. In agent systems, each agent considered to have their own thread of control whereas single thread on control in objects. And this review finds the definition of multi agent system and the issues in future direction.

Sharjeel Imtiaz et al (2005) presented a framework for unification of information extraction and data mining. This framework is designed by an intelligent agent. The intelligent agent is designed to extract features with cross feedback approach and also provides unified undirected graphical handle. It creates a good model for both entity and text level abstraction.

Borowski & Hopkinson (2011) explored integrating a reputation-based trust mechanism with an agent-based backup protection system to help and to protect against complex network failures. The goal of this research is to
have the ability to find and react to malicious behavior while remaining compatible with traditional network security elements such as firewalls and data encryption. To achieve this goal, a reputation-based trust system has been integrated with an intelligent agent-based backup protection system, enabling agents to analyze other agents’ behavior. The success of this approach is determined by a performance comparison with the original agent-based backup protection scheme as well as with traditional protection mechanisms when malicious behavior interferes with communications between the communication-based agents.

Yaskawa & Sakata (2003) presents an application of intelligent agent technologies to different real world problem such as semiconductor manufacturing process, train operations and electric motor assembly lines. It breaks the conventional simulators which has limited values for examination. In this agent construction, an additional platform Parallel Inference Machine has been used that describes and executes multiple agents in the independent and parallel manner.

Don Gilbert (1997) explored the value of agents by examining some real agent-enhanced applications. This paper explains the applications of intelligent agent like customer helpdesk, web browser intelligence, shopping assistance etc. It lists the basic characteristics of intelligent agent. Finally the agent architecture of IBM named Agent Building Environment (ABE) is introduced with the basic functions.

2.3 WEB MINING

Etzioni (1996) has proposed the web techniques to automatically discover and extract information from Web documents and services. This area of research is so huge and the interest is getting increased because of different research communities, the tremendous growth of information sources on the
Internet and the recent interest in e-commerce. Web mining field consists of three main categories, web usage mining, web structure mining and Web content mining. Web usage mining refers to the discovery of user access patterns from web usage logs. Web structure mining tries to discover useful knowledge about the structure of hyperlinks. Web content mining aims to extract or reduce useful information or knowledge of website content. Web mining is often associated with information retrieval and information extraction. However, Web mining is not the same as IR or IE.

2.3.1 Information Retrieval

Sandor Dominich (2000) proposed IR technique which is the automatic retrieval of all relevant documents, while at the same time as some of retrieving irrelevant as possible. Some have claimed that resource or document discovery on the Web an instance of web content mining and web mining other employees with intelligent IR. Actually, the primary goals of indexing text and searching for useful documents in a collection and research today in IR includes modeling, document classification and categorization, user interfaces, data visualization and filtering.

Markov et al (2006) proposed the task that can be considered to be an example of web mining, web document categorization or classification which could be used for indexing. Viewing in this respect, web mining is part of the web IR process.

2.3.2 Information Extraction

Cowie & Lehnert (1996) proposed an information extraction is the goal of transforming a set of documents, usually with the help of an IR system, into information that is more readily digested and analyzed. IE aims to extract relevant facts from the documents while IR aims to select relevant
documents. While IE is interested in the structure or representation of a
document, IR views the text in a document just as a bag of unordered words.
Thus, in general IE works at a finer granularity level than IR dose on the
documents level.

El-Beltagy et al (2007) proposed an example of IE without web
mining for building a model for automatically augmenting segment
documents with metadata using dynamically acquired background domain
knowledge in order to assist users in easily locating information within these
documents through a structured front end.

Guowei Chen & Pengzhou Zhang (2012) presented a method of
transforming web unstructured information into structured information based
on knowledge base. Because knowledge base supports the ordering of
information and knowledge. It will speed up the knowledge and the flow of
information and make for knowledge sharing and communication. The
information extraction based on wrapper, statistic and data mining is being
used to extract webpage information.

Yaqing Liu et al (2009) proposed a hierarchical extraction model
based on Hidden Markov model. It is considered as the best model when
compared with some existing algorithms used in the field of web information
extraction. The annotated information of atom information items and
compound information items are used with the compound information items’
paths in DOM+ (Document Object Model) tree to build the hierarchical
extraction model.

Information extraction tools have been exploited to help users to
automatically find required information in web pages. Those tools can be
classified into two types: one is based on special grammars and others based
on rule induction. Firstly, these tools use a particular kind of query language
to express user’s requirements and then use an interpreter to extract the
information by query expression. Obviously, extracting information by using
the tools based on special grammars depends on the path of the location of the
information. It will implicitly impose mastering particular query language on
users. The methods based on rule induction lay keystone of how to
automatically generate wrappers for these web pages with similar structure.

Luis Tari et al (2010) developed information extraction model by
using database queries. The parse trees are used here to store the output by
text processing in a database, and then express extraction needs using queries,
which can be evaluated and optimized by databases. Database queries for
information extraction enable generic extraction and minimize reprocessing
compared with the existing approaches. However, such an approach also
poses a lot of technical challenges such as language design, optimization and
automatic query generation.

Using this approach, a general-purpose information extraction
system, GenerIE is prepared in the context of Bio-medical extraction. It can
efficiently manipulate diverse extraction needs and maintain the extracted
information up-to-date incrementally when new information becomes
available. This research paper discusses the main advantages of this GenerIE
framework

First, using database queries instead of writing complicated special-
purpose programs, information extraction becomes generic for diverse
applications and becomes very easier for the user. The user can edit and
analyze an extraction pattern by writing a database query. Next, with the use
of databases, GenerIE only needs to perform extraction incrementally on the
information that are affected by an revised module, and thus it is much more
efficient than running the whole extraction programming code from scratch as
required by existing systems.
Robert Novonty et al (2009) presented a chain of techniques for object attribute data extraction from web pages which contain either multiple object data or detailed data about a single object. The object attribute values extraction mainly used for two kinds of web pages, namely master pages containing structured data about multiple objects and detail pages, which contain data about single product.

The traditional web information extraction is mainly directed based on static HTML pages. To extract information in dynamic web pages in an effective way becomes one of the difficult issues in the information extraction field. Zhang Hengru & Cui Chun (2011) developed a method about web information extraction based on Ajax through the thorough investigation into the traditional web information extraction. The Ajax is a technique which defines asynchronous characteristic of the user operations and the server response. Compared with the traditional web information extraction technology, its advantage is embedded with a dynamic web processing engine in the process of DOM (Document Object Model) tree procedure. This method will be used exclusively to analyze the dynamic web pages for the DOM tree and in real-time, update the DOM tree.

Sunita Sarawagi (2008) presented that manual building IE system is not feasible and scalable for a dynamic and diverse medium such as web contents. Due to this nature of the web, most IE systems focus on specific web sites to extract. Others use machine learning or data mining techniques to learn the extraction patterns or rules for web documents semi automatically or automatically. Within this view, web mining is used to improve web IE (web mining is part of IE).

Qian Zhu & Xianyi Cheng (2008) presented the challenges and opportunities of Information Extraction. Both challenges and opportunities are mainly discussed in three fields namely Knowledge management, E-
Commerce and Question Answering System. In the knowledge management concern, to organize knowledge, the concept space is used according to the meaning of text. The knowledge maintenance is improved by automatic software that has the function of compatibility check. The knowledge management must support Question Answering on cross-document.

It provides a survey of information extraction research of over two decades from different communities of researchers bringing in techniques from machine learning, databases, information retrieval, and computational linguistics. Different dimensions are derived for review from the nature of the extraction task, the techniques used for extraction, the variety of input resources exploited, and the type of output produced. The rule-based and statistical methods for entity and relationship extraction is elaborated and surveyed techniques for optimizing the various steps in an information extraction pipeline, adapting to dynamic data, integrating with existing entities and handling uncertainty in the extraction process.

### 2.4 TEXT MINING

The related work of information extraction aims to find specific data in natural language text. The rapidly evolving field of text mining has seen advances in the agent based tools available for text mining. It builds largely on research on Data Mining.

#### 2.4.1 Text Mining for the Web Documents

As discussed earlier, text mining is often considered as a sub-field of data mining and refers to the extraction of knowledge from text documents (Chen 2001). Because the majority of documents on the Web are text documents. Text mining for Web documents can be considered as a sub-field of Web mining, or, more specifically, Web content mining. Information
extraction, text classification, and text clustering are examples of text-mining applications that have been applied to Web documents.

Although information extraction techniques have been applied to plain text documents, extracting information from HTML web pages can present a quite different problem. However, web pages are also comparatively unstructured. Instead of a document consisting of paragraphs, a web page can be a document composed of a side bar with navigation links, tables with textual and numerical data, capitalized sentences, and repetitive words. The range of formats and structures are very diverse across the Web. If a system could parse and understand such structures, it would effectively acquire additional information for each piece of text. For example, a set of links with a heading “Link to my friends’ homepages” may indicate a set of people’s names and corresponding personal home page links. The header row of a table can also provide additional information about the text in the table cells. On the other hand, if these tags are not processed correctly but simply stripped off, the document may become much noisier.

Kin Keung Lai (2006) developed a multi-agent web text mining system on the grid to support enterprise decision making. It used a framework of the Back Propagation Neural Network (BPNN) based intelligent learning agent for text mining. It refers the process of using unstructured web-type text and examining it to discover implicit patterns hidden within the web documents. The BPNN is used as an intelligent agent to explore the hidden patterns. BPNN agent is a supervised learning mechanism in the form of the neural network associative memory. But single BPNN agent could not handle large-scale text documents. With the rapid increase of web information, a multi-agent web text mining system on the grid is then constructed for large-scale text mining application. The multi-agent based web text mining on the
grid can discover some useful knowledge for enterprise decision support in an efficient way.

Text mining is implemented through Entity – Relationship based on Information extraction and presented by Lipika Dey et al (2007). Text mining identifies key information components and mined from the repository are chained using an N-gram-based algorithm. This information chains provide a comprehensive view of the collection of the document. Summarizing a document through relations help in easy visualization of contents of a repository and experimental Chaining through relations can also help in the discovery of potentially interesting information from the vast text repository.

Chang & Lui (2001) used a PAT tree to construct automatically a set of rules for information extraction. The system, called IEPAD (Information Extraction Based on Pattern Discovery), reads an input Web Mining: Web page and looks for repetitive HTML markup patterns. After unwanted patterns have been filtered out, each pattern is used to form an extraction rule in regular expression. IEPAD has been tested in an experiment to extract search results from different search engines and achieved a high retrieval rate and accuracy.

Christina Feilmayr (2011) presented an extended methodology for developing text mining supported Information Extraction systems. The proposed work integrates data mining with information extraction methods. It provides a new high-quality information extraction methodology and, at the same time, to improve the performance of the underlying extraction system. Consequently, the proposed methodology should shorten the life cycle of information extraction engineering because information predicted in early extraction phases can be used in further extraction steps, and the extraction rules developed require fewer arduous test-and-debug iterations. Effectiveness
and applicability are validated by processing online documents from the areas of e-Health and e-Recruitment.

Ning Zhong et al (2012) presented an innovative and effective pattern discovery technique which includes the processes pattern evolving and of pattern deploying, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Low frequency and misinterpretation are two fundamental issues regarding the effectiveness of pattern-based approaches. In order to solve these issues, an effective pattern discovery technique is developed, which first calculates discovered specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns rather than the distribution in documents for solving the misinterpretation problem. It also considers the influence of patterns from the negative training examples to find ambiguous patterns and try to reduce their influence for the low-frequency problem.

2.4.2 Structured Data Extraction

This is perhaps the most widely studied research topic of web content mining. One of the reasons for its importance and popularity is these structured data on the web are often very important as they represent their host pages. Essential information is like lists of products and services. Extracting such data allows one to provide value added services, like comparative shopping, and meta-search. Structured data are also easier to extract compared to unstructured texts.

This problem has been studied by researchers in AI, database and data mining, and web communities. There are several approaches to structure data extraction, which is also called wrapper generation.
The first approach is to write manually an extraction program for each website based on observed format patterns on the site. This approach is very labor intensive and time-consuming. It thus does not scale too many sites. The second approach is wrapper induction or wrapper learning, which is the main technique currently. Wrapper learning works as follows: The user first manually labels a set of trained pages. A learning system then generates rules from the training pages. The resulting rules are then applied to extract target items from web pages. Since structured data objects on the web are normally database records retrieved from underlying databases and displayed in web pages with some fixed templates. Automatic methods aim to find patterns or grammars from the web pages and then use them to extract data.

2.4.3 Unstructured Data Extraction

Philipp Cimiano et al (2005) proposed use of common language patterns and redundancy of information on the web to find concepts, relations among concepts and named entities. Common sentence structures used to express certain facts or relations.

Abidin et al (2010) have proposed patterns that are automatically learned or supplied by users. Another direction of research in this area is web question-answering. Although question-answering was first studied in information retrieval literature, it becomes very important on the web as the web offers the largest source of information and the objectives of many web search queries are to obtain answers to some simple questions and extends question-answering to the web by query transformation, query expansion, and then selection.

Lu (2010) had proposed that the most web pages can be seen as text documents. Extracting information from web documents has also been studied by many researchers. The research is closely related to text mining,
information retrieval and natural language processing. Current techniques are mainly based on machine learning and natural language processing to learn extraction rules from manual labeled examples.

2.5 INTELLIGENT AGENT METHODOLOGY

Intelligent agent provides a methodology to realize an autonomous decentralized system with cooperative interactions among agents that model each element of the system. It has problem solving and learning skills, as well as the knowledge for that purpose. Many research works in implementation of real world problem solving and simulation model have been carried out by using agent methodology.

2.5.1 Intelligent Agent Framework

Intelligent agent is implemented to equalize the human behavior and believes. Nicole Ronald & Leon Sterling (2005) designed the agent based model of pedestrian behavior in a real world environment. It defines the unpredictable nature of human decision making. Human decision exhibits different behaviors depending on their knowledge on that environment and other personal characters. BDI (Belief – Desire – Intention) agent based architecture model is implemented for this simulation which is useful for high level decision making. The capable of doing several things concurrently without trouble is considered as one more additional advantage of agent.

Ji Ma & Mehmet Orgun (2009) proposed a formal method to establish trust theories for communication protocols in agent based systems using temporal belief logic. The authentication protocol is defined and used for trust management between agents. It is used to verify and authorize agents acting on behalf of users to protect restricted information. After
authentication, two agents should be entitled to believe that they are communicating with each other and not with intruders.

The research and commercial community has primarily focused on designing intelligent agent in specific domain knowledge with embedding rules and a trained data. Using the trained data and the intelligent agent, the information retrieval will provide flexible, modular and delegated solution.

Raymond Lee (2004) presented an integrated framework for data retrieval and data extraction in the context of Internet shopping. To computerize a series of product search and selection activities, the work focuses on implementing agent technology with mining. It is based on the intelligent agent which is implemented by JADE - Java Agent DEvelopment Environment which is a multi agent platform. The framework mainly concentrates on automatic fuzzification and defuzzification scheme.

Maja Pantic (2005) presented a flexible technique for teaching introductory artificial intelligence using java and implemented simple agent framework. It uses intelligent agent and other Artificial Intelligence techniques to mention filter and retrieve relevant information from the World Wide Web.

Giuseppe Calafiore et al (2012) proposed the measurement of inter-agent distance and communication of each agent in a distributed environment. The underlying geometrical problem has been studied quite extensively in different fields like chemical engineering, bio technology and robotics, and it is known to lead to a hard nonconvex optimization problem. It is mainly focused for computing the position of each agent automatically, communicating the data within the neighbor without the help of central station and without the knowledge of structure of the network.
One of the main applications of agent is network security. Dasgupta et al (2005) describes the security agent architecture called CIDS which is used as an administrative tool for intrusion detection. It allows easy inclusion of new detection, decision and action plug-in, independently. It can also act as monitoring tool for data gathering and visualization which help to evaluate the behavior of any monitored network.

Bathy (2011) described an Intelligent Extended Clustering Genetic Algorithm (IECGA) using Business Process Execution Language (BPEL) to make best possible solution for data clustering. It advanced the efficiency and performance for recovering suitable information and the results which satisfied the user’s requirements. In data clustering, IECGA gives the required document for the user, based on the similarity amid query matching and relevant document.

2.5.2 Multi Agent Systems Framework

Vallejo et al (2010) presented a multi-agent framework for facilitated and administrated work for agent based application. JADE – Java Agent Development Environment tool is proposed a programming tool implemented by java for the development of multi agent application. But it has the restriction of low efficiency and programming language restriction. That has been improved by improving scalability and short time response of the tool. This is achieved by object oriented methodology.

In the multi-agent platform, The FIPA (Foundation for Intelligent Physical Agent) services are automatically deployed which are highly configurable depends on the application. It can be integrated with web administrative system to remotely manage multi-agent application.
Sheng-Yuan Yang & Yi-Yen Chang (2010) proposed network management system by ontology supported multi agent techniques. It is mainly designed for reducing recovery time of network troubleshooting in the network management. This system is integrated with the intelligent agent framework to effectively enhance and strengthen the network monitoring performance, and accordingly produce related quantification values of dynamic information. This system obtained the information from the coordination and cooperation of multi-agent system. The ontological free software is used for storing the operating information on network management in backend database. The system is obtained its adoptability from the open source code which makes the user can able to control irregular phenomena in the network.

2.5.3 Agent Based Tools

Mehdi Dastani (2008) presented a BDI-based agent-oriented programming language, named 2APL -A Practical Agent Programming Language. BDI is one of the agent architecture characterized by ‘mental state’ with three components called belief, desire and intention. Belief is information that an agent has about the environment, desire is an option available for agent to take a decision and intention represents state of affairs that the agent has chosen and committed resources to. This language provides an effective integration of declarative and imperative style agent programming by introducing declarative beliefs and goals with events and plans.

The agent-oriented software engineering framework provides a cognitive concepts and abstractions in terms of the way of software model which can be specified, designed, and implemented. Because most agent-oriented analysis and design methodologies assist system developers to
specify and design system architectures in terms of agent concepts and abstractions.

This programming language is designed with the following characteristics.

- It defines and integrates programming constructs that are expressive enough to implement a variety of agent concepts and abstractions used in the existing agent-oriented programming methodologies.

- It creates an interface between the expressivity of programming constructs designed to represent and reason with agent concepts, and the expressivity of constructs designed to implement the update and revision of those concepts.

- It has formal semantics such that it is possible to verify whether agent programs satisfy their formal specifications.

- It captures important and intuitive rationality principles of agent concepts.

- It realizes an effective integration of declarative and imperative programming styles.

Generally multi agent system can be implemented in any programming language. However this language is designed to provide dedicated and expressive programming constructs to facilitate practical and effective implementation of agent related concepts.

One of the major innovative technologies of agent is development of distributed software system. Fabio Bellifemine (2008) presented a framework JADE – Java Agent Development environment that facilitates
development of interoperable intelligent multi-agent systems. This paper presents JADE and its technological components together with a discussion of the possible reasons for multi agent systems.

JADE is a middleware that enables fast and reliable implementation of multi-agent distributed systems and which can be integrated with Artificial Intelligence (AI) tools. JADE is fully developed by Java with the principles of interoperability, uniformity and portability. The comparison of JADE with other toolkit has been done and the JADE is leading with the following features. Firstly JADE is completely based upon a FIPA specification. Then, JADE provides a proper set of functions to support development of multi-agent systems but it puts very few restrictions on the user for coding.

Users must simply use and write Java code without learning any new special construct. Finally, JADE can be deployed on JEE, JSE, and JME devices and it provides a homogeneous set of APIs that is independent of any network and Java technology. The system architecture and the specialization of agent for mobile environment are discussed.

![Figure 2.1 The architecture of JADE agent system](image-url)
The JADE architecture is completely designed based on peer-to-peer ideas and communication patterns. In this architecture, the intelligence, the information and the control are fully distributed across a group of heterogeneous hosts, including mobile terminals and PDAs. JADE components can be categorized as agent and services. Services are operations triggered by agents. Each instance of the JADE run-time is called container which contains agents. The set of all containers is called platform and provides a homogeneous layer that hides to agents. Many companies have used JADE for network management, supply chain management, rescue management, fleet management, healthcare, auction and tourism etc.

Fabio Bellifemine et al (2001) presented JADE, a software framework to write agent applications in compliance with the FIPA (Foundation for Intelligent Physical Agents) specifications for interoperable intelligent multi-agent systems. It introduces FIPA specification and main features of JADE. It also describes the architecture of the agent platform, the communication subsystem.

The Foundation for Intelligent Physical Agents (FIPA) is an international nonprofit association sharing the effort to produce specifications for generic agent technologies. FIPA does not just promote a technology for a single application domain but a set of general technologies for many application areas that developers can integrate to make complex systems with a high degree of interoperability.

2.6 AGENT BASED INFORMATION EXTRACTION

Fuchun Peng & Andrew McCallum (2006) presented Conditional random fields (CRFs) for extracting various common fields from the headers and citation of research papers from online. To address the constraint information extraction issue, an approach for creating canonical citations for a
publication and improving citation segmentation is proposed. It also defines co-reference constraint information extraction.

Pericles Mitkas et al (2007) developed Agent Academy (AA), a software platform for the design, creation, and deployment of Multi Agent System (MAS), which combines the power of knowledge discovery algorithms with the versatility of agents to control a large number of Concurrent Engineering (CE) task. The intelligence can range from rudimentary sensor monitoring and data reporting, to more advanced forms of decision-making and autonomous behavior. The behavior of each agent can be obtained by performing data mining on the available data and the respected knowledge domain. It mainly pinpointed the inherent correlation between CE and agent technology. This architecture provides the capability of grasping and reproducing the core prerequisites of the CE design processes. It provides high-level design capabilities and deal with the internals of the agent architecture, in order to be considered complete and generic.

Raymond Mooney & Un Yong Nahm (2003) presented a text mining framework using a learned information extraction method to transform text into more structured data which is then mined for important relationships. Rules mined from a database extracted from a corpus of texts are used to predict additional information to extract documents, thereby improving the recall of the underlying extraction system.

The experiments are tested and showing that rules discovered from an agent extracted database are close in accuracy to that discovered from a manually constructed database. The initial task of the framework is designing a database by applying a learned information-extraction system to a natural-language document. Then, it applies standard data-mining techniques to the extracted data, discovering knowledge that can be used for many tasks, including improving the accuracy of information extraction. The rules are
induced for predicting each piece of information in each database field given all other information in a record. It enables the application of KDD (Knowledge Discovery from Database) to unstructured text and KDD can discover predictive rules useful for improving IE performance. It is finally concluded that text mining is a relatively new research area at the intersection of natural-language processing, machine learning, data mining, and information retrieval.

Hratch Mangassarian & Hassan Artail (2007) presented an information extraction strategy for handling subjective information from unstructured text. The described framework comprises four sequential processing steps: part-of-speech tagging, syntactic parsing, relation generation and criteria evaluation. This technique assesses the significance of online news about a company on its stock price. It could be integrated in a larger system that would monitor news and company sites on the Internet. As soon as matching news would break out, it would instantaneously and automatically analyze it, present the important information to a user and in extreme cases automatically set off buying or selling stocks. The basic idea is to compile a set of preferably independent criteria that are sufficient to evaluate the given entity such as company, job application, essay and related task in the considered application field. This set of criteria and their corresponding relations should preferably be compiled by field experts.

Jae-Woo Lee (2007) proposed a simple model for information retrieval agents based on many terms or keywords distribution in a document or distributed database. The techniques meaningful term’s frequency and the key word distribution characteristics are used in this model for the key paragraph extraction in a document. The terms are selected by using stemming, filtering stop lists, synonym for retrieving meaningful terms in a document. The designed agent receives a web client’s information retrieval
request and extracts key paragraph with frequency and distribution using the keywords of the client, and then the agent constructs profile of the documents with the keywords, key paragraph, and address of the document browsing.

Giuseppe Dell Penna et al (2010) presented a graphical software system that provides an automatic support to the extraction of information from web pages which exploits the visual appearance of the information in the document. This technique has been integrated within the Spatial Relation Query (SRQ) tool which allows one to define and manage a library of spatial relations.

Ning Zhong & Shinichi Motomura (2009) described an agent-enriched mining approach demonstrates the Brain Informatics (BI) system. This can be obtained by transforming and mining human-brain data from cognitive event related potential experiments. It understands human Information Processing (IP) principles and mechanisms relating to higher cognitive functions such as problem solving, reasoning, and learning.

Longbing Cao et al (2007) present overview of the agent-mining interaction from the perspective of an emerging area in the scientific family. It summarizes key driving forces, originality, major research directions and respective topics, and the progression of research groups, publications and activities of agent-mining interaction. It addresses both theoretical and application-oriented aspects. It discusses the research directions in Agent-Mining interaction and mutual enhancement issues in Agent-Mining interaction and application of this interaction.

Kyriakos Chatzidimitriou & Andreas L. Symeonidis (2009) explained that any technology coupled with agent will provide real-time input and assessment of the data mining extracted models and decisions on the current state of the environment can be made in an automated manner.
Specifically, in a supply-chain organization, data mining can be applied to various facts such as pricing, forecasting, and customer and supplier relationship management, always keeping in mind that these solutions should satisfy all security, safety, and soundness issues that might arise in such versatile environments.

Andreas Symeonidis et al (2007) presented an approach that provides a gateway on the way data mining techniques can be employed in order to augment agent intelligence. It demonstrates how the extracted knowledge can be used for the formulation, and the improvement, in the long run, of agent reasoning. It is identified that data mining in association with agent-based applications offers a strong autonomous and adaptable framework. Three dominant data mining techniques such as clustering grouping, classification categorization and prediction, and association rule extraction for correlation discovery are applied to Multi-agent system for incorporating resulting knowledge into agent reasoning. This approach takes under serious consideration the need for knowledge model evaluation and provides a series of functionalities for visualization, model testing and model comprehension.

Chia-Hui Chang et al (2003) proposed a pattern discovery approach to the rapid generation of information extractors that can extract structured data from semi-structured Web documents. It applies several pattern discovery techniques, including PAT-trees, multiple string alignments and pattern matching algorithms that discovers extraction patterns from web pages. The framework is designed with three components namely pattern discoverer which accepts an input Web page and discovers potential patterns, Rule generator which remembers the pattern and save it as a extraction rule for later applications and Extractor extracts desired information from similar web pages based on the designated extraction rule. It also provides a
multilevel alignment technique to extract finer data from discovered data records. The limitation of this approach is also identified that it cannot be applied to web pages that contain only one data record and the extraction rule for one web data source generalized poorly to other web data sources with different layout formats.

Velasquez & Vasile Palade (2007) introduced a Knowledge Base (KB), which has a database-type repository for maintaining the patterns, and rules, as an independent program that defines the pattern repository. In order to increase the relation between the web site and its visitors an artificial system or a human user can consult the KB by using this architecture. It was defined by rules and the knowledge repository. It was tested with the bank data and the effectiveness proved. Before applying web mining techniques, the data are transformed into behavior patterns, using a particular model about the visitor behavior.

It proposed to maintain the patterns by storing them in a database-like repository and the rules as an independent program that consults the patterns repository when preparing the recommendations. The representation of knowledge as a set of rules is not the best approach, as the number of rules could increase in the time, therefore making difficult on maintenance. The proposed KB deals with this instance by maintaining patterns and parametric rules. This framework provides a minimum set of rules, an easy maintenance of them and a significant repository of patterns.

The information retrieved in common search engines provide too much and unstructured information to the user. Eddie Chan (2009) proposed an agent-based technique for extracting location aware data that can improve the speed of extraction while maintaining or even improving the accuracy by making use of semantic information in the data to develop smaller training
sets. This framework classifies, matches and organizes the information from the WordNet English dictionary to find a best match for a user query.

It provides number of benefits. First, the cognitive semantics graph model provides a hierarchical structure to the location-aware data. Second, it prevents the need for extensive manual information grasping. Third, the modified classical information extraction improves the accuracy of matching keywords and provides a more cognitive semantics meaning result. Finally, the agent can update the information and directly establish a communication with its neighbor agents. The high classification technique TFTIDF (Term – Frequency – Time – Inverse - Document - Frequency) is used for query processing. It formulates the significance of a term according to its frequency in documents. TFTIDF is being used as a weight of term in document. Weight of the term $W_{di}$ is calculated as follows. If a term $t$ occurs in document $d$,

$$W_{di} = tf_{di} \times \log (N / idf_{di})$$ \hspace{1cm} (2.1)

where

t, is a term in document collection

$W_{di}$ is the weight of $t$

$tf_{di}$ is term frequency (term count of each word in a document) of $t$

$N$ is the number of total documents in the collection

$idf_{di}$ is the number of document in which $t$ appears

Four different agents are used for grasping, classification, matching and organizing the data. Information gathering agent gathers the information automatically from different information source and then updates the information from the neighbor agent or other system.
This agent always obtains the latest information throughout the system. The classification agent will classify output information into several categories. It calculates the Euclidean distance between the words, and it will form a word graph which contains the words relationship. Information matching agent is designed to match the user query and the categorized information. Information organizing agent is to organize the search result based on the user’s preference. It provides the user-friendly interface to user.

The categorization effectiveness of the performance of precision rate and speed are compared with other two techniques, three layer neural network and TFID with hierarchical clustering. The precision rate shows 96.68% which is highest among three techniques. The computational time taken for finding suitable information in this technique is low compared with other two techniques.

Ramesh Thakur et al (2012) proposed a novel and hybrid approach of learning context-free grammar rules that are based on alignment between text documents. Also it automatically discovers the grammar rules using grammatical inference of repeated pattern present in un-structured text document. These rules can be used to infer the attribute value pairs from the unstructured text document. It mainly extracted attribute value of frequently occurring data from data intensive document.

Mohammed Kayed & Chia-Hui Chang (2010) formulated the data extraction problem as the decoding of page generation based on structured data and tree templates. For that an unsupervised, page-level data extraction approach is proposed to deduce the schema and templates for each individual Website. It applies tree matching, tree alignment, and mining techniques to achieve the task. The page generation model is formulated using an encoding scheme based on tree templates and schema, which organize data by their
parent node in the DOM trees which can be recognized and merged into the pattern tree for schema detection.

Ying Chen et al (2010) proposed system which extracts lightweight features with a totally unsupervised approach from broad resources for constructing a social network like semantic web. To extract the information of the focus person, an integrated system is introduced, which is able to effectively re-use and combine current well-developed tools for web data. Personal named entity ambiguity and the information extraction for a specific person are considered as basic issues for construction of semantic web. This system extracts features with various lightweight methods and from broad resources, such as downloading related web pages, using a large web corpus, and extracting the information from a search engine.

Yevgen Biletskiy et al (2009) presented an application for creating a database by automatically extracting relevant information from HTML course outlines stored on an institution’s website and storing it in a local machine as machine-readable XML form for the proposed semi-automatic academic e-Advisor. It parses a course outline based on its HTML tags and content to build a Document Object Model then applies a combination of web mining, natural language processing, and pattern recognition techniques.

Hanmin Jung et al (2005) introduced the techniques verified on the question answering framework, such as domain knowledge and instance rules, into an information extraction problem. To improve extraction performance, a sequence of the user-oriented learning and the separate-context learning produces context rules and generalizes them in both the learning and extraction phases.

The main goal of the system is to develop a system that retrieves answers rather than documents in response to a question. The IE model on a
question answering framework improves the building process of domain knowledge by separately applying the types of web documents. This technique is designed as three phases: building which constructs several classes of extraction knowledge, learning which generalizes the rules to enhance the extraction coverage, and extraction which extract target frames using the knowledge obtained and generalized by the building and learning phase. It uses structured Web documents, dictionaries, and semantic information to seed patterns of instances and context.

A cognitive semantic recovery system based on agent is proposed by Eddie Chan (2009) for location aware data. Here the intelligent agent is used for the purpose of data grasping, categorization matching and organizing. It initially collects location aware data and use semantic graph in the world net English dictionary. They categorize, match and organize the data for finding a best match for a user query.

Raymond Mooney & Razvan Bunescu (2003) discussed two approaches to using natural language information extraction for text mining. First approach is extraction of general knowledge directly from text and the second approach is extraction of structured data from text documents or web pages and then applies traditional KDD methods to discover patterns in the extracted data.

2.7 CONCLUSION

In this chapter the related work of this research work and existing techniques for Information Extraction by agent framework and the fields of data mining and text mining are discussed. It is observed that Information extraction is related to Data Mining and text mining where text files and databases are preprocessed to improve subsequent mining tasks by filtering irrelevant or useless information.