OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Data and Information Integration: Information Extraction

Varnica Verma[1]

[1](Department of Computer Science Engineering, Guru Nanak Dev University, Gurdaspur Campus, Punjab, India)

ABSTRACT: *Information extraction is generally concerned with the location of different items in any document, may be textual or web document. This paper is concerned with the methodologies and applications of information extraction. The field of information extraction plays a very important role in the natural language processing community. The architecture of information extraction system which acts as the base for all languages and fields is also discussed along with its different components. Information is hidden in the large volume of web pages and thus it is necessary to extract useful information from the web content, called Information Extraction. In information extraction, given a sequence of instances, we identify and pull out a sub-sequence of the input that represents information we are interested in.*

Manual data extraction from semi supervised web pages is a difficult task. This paper focuses on study of various data extraction techniques and also some web data extraction techniques. In the past years, there was a rapid expansion of activities in the information extraction area. Many methods have been proposed for automating the process of extraction. We will survey various web data extraction tools. Several real-world applications of information extraction will be introduced. What role information extraction plays in different fields is discussed in these applications. Current challenges being faced by the available information extraction techniques are briefly discussed along with the future work going on using the current researches is discussed.

Keywords: *DELA, HTML, IE, NLP, REES.*

I. Introduction

Information Extraction (IE) is used to identify a predefined set of concepts in a specific domain, ignoring other irrelevant information, where a domain consists of a corpus of texts together with a clearly specified information need. In other words, IE is about deriving structured factual information from unstructured text. For instance, consider as an example the extraction of information on violent events from online news, where one is interested in identifying the main actors of the event, its location and number of people affected.

Information extraction identifies class of pre-specified entities and relationships and their relevant properties. The main aim of information extraction is to represent the data in the database in structured view i.e., in a machine understandable form [6].

II. Literature Survey

According to Jakub Piskorski and Roman Yangarber [1], information extraction is an area of natural language processing that deals with finding information from free text. Sunita Sarawagi [2] studies the different techniques used for information extraction, the different input resources used and the type of output produced. She says that information extraction can be studied in diverse communities. Devika K and Subu Surendran [3] provide different tools for web data extraction. Jie Tang [5] details on the challenges in the field of information extraction.

2.1 Early Years: Knowledge Extraction Systems

In the early years, information extraction systems were developed using the knowledge engineering approaches where the creation of knowledge in the form of rules and patterns, for detecting and extracting the required information from the database, was done by human experts. Most of the early IE systems had a drawback that they showed a black-box character, which were not easily adaptable to the new scenarios. The aim of knowledge based systems is put efforts for general purpose information extraction systems and frameworks which are easier to adapt and learn by the new domains and languages [1]. Modularized approach was used for the development of such systems. Two examples of modular approach used are IE2 and REES. The first one achieved the highest scores for all IE tasks and the second had been the first attempt for large scale events and relation extraction systems based on shallow text analysis methods.

2.2 Architecture: Components of Information Extraction System

Different IE systems are developed for performing different tasks but some components always remain the same. These components typically include core linguistic components-which are useful to perform NLP (natural language processing) tasks in general- and IE specific components which address the IE specific tasks. Also, domain-independent and domain specific components are also included [1].

The following steps are followed in order to extract information:-

2.2.1 Domain Independent Components

- Meta-Data Analysis: - Extracts the title, body, structure of the body and the date of the document i.e., the date when the document was created [1].
- Tokenization: - Text is segmented into different parts called tokens which constitute words with capital letters, punctuation marks, numbers used in the whole text etc. All the tokens are provided with different headings and the relevant data from the text is added to the respective tokens [1].
- Morphological Analysis: - In this step the information is extracted from the tokens [1].
- Sentence or Utterance Boundary Detection: - This performs the formation of sequence of sentences from the text, along with different items associated with it and their different features [1].
- Common Named-Entity Extraction: - Extraction of domain-independent entities is performed. These entities are represented with common names like number, currency, geographical references etc [1].
- Phrase Recognition: - Recognition of verbs, nouns, abbreviations, prepositional phrases etc. is performed in this step [1].
- Syntactic Analysis:-Structure for sentences is designed based on the sequence of different items being used in the sentence. The structure can be of two types: one is deep i.e. includes each and every detail of the items being consumed and second is shallow i.e. includes only the specific items and not their further properties or attributes. The structure can be like parse trees. The shallow structure fails to represent ambiguities (if any) [1].

2.2.2 Domain Specific Components

The core IE tasks are domain specific and are therefore implemented by domain specific components. Domain specific tasks can also be performed at lower level of database extraction. The following steps are applied:-

- Specific Named-Entity Recognition: - The text is extracted using some specifically highlighted terms from the text. For example, in domains related to medicine some specialized terms related to medicine are required whereas in case of a large enterprise there is no such requirement [1].
- Pattern Matching: - The entities and their key attributes are extracted. These must be relevant to the target relation or event. All the properties of each and every entity are detected from the text. The entities constitute different patterns according to the properties inherited by them [1].
- Co-Reference Resolution: - Implementation of inference rules is done in this step in order to create fully fledged relations or events [1].
- Information Fusion: - The entities with same attributes are combined to constitute one entity set. The related information is generally grouped together in different sentences and documents. All this data is collected and grouped together in the pattern of their properties so that a proper relation can be made [1].

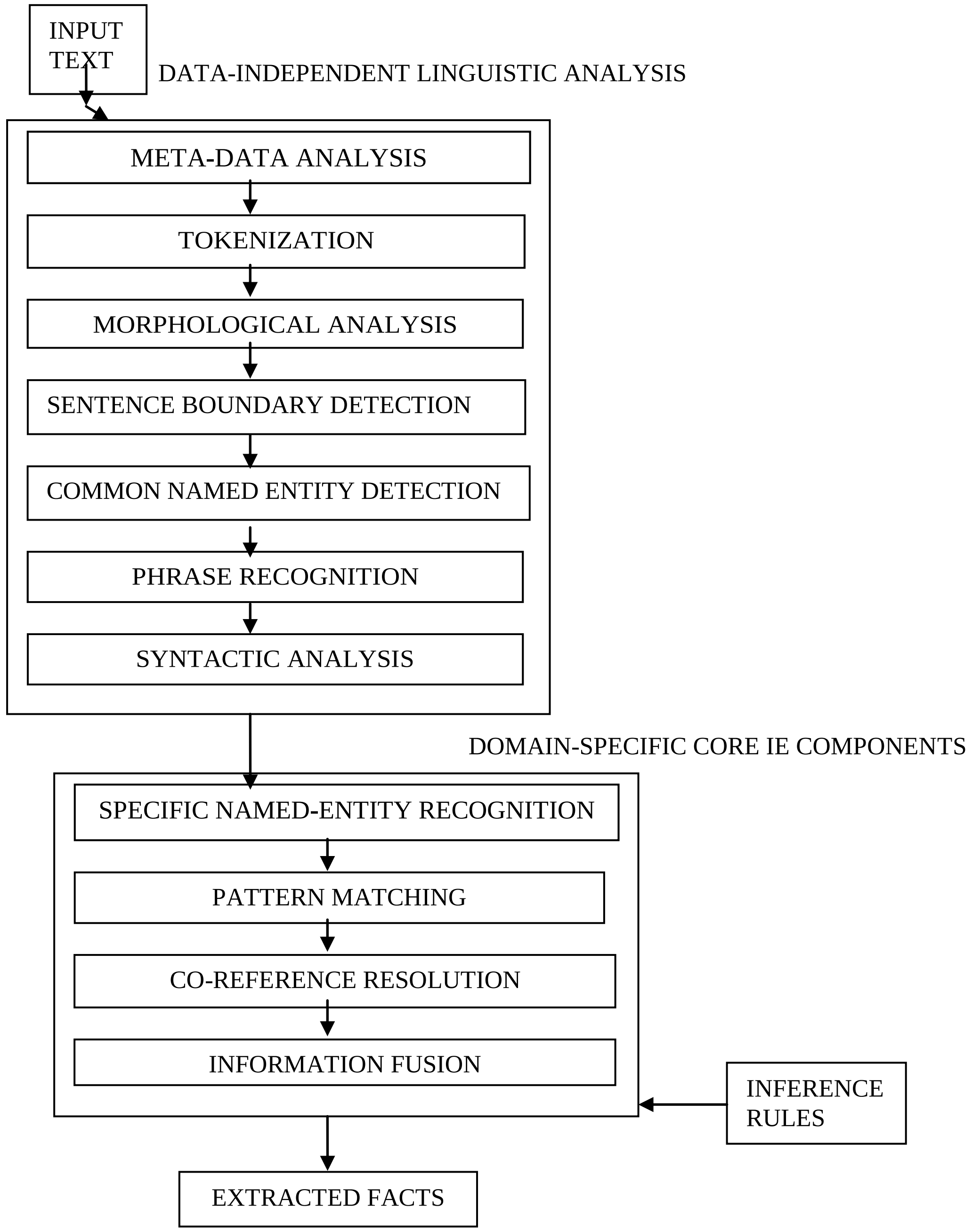

Figure: 1. Architecture of information extraction system

2.3 Knowledge Extraction Techniques

2.3.1 Rule Based Technique

The detection and extraction of information and data is performed by using some knowledge based rules. Human expertise plays an important role in this method [5].

Advantages:

- Fast
- Simple
- Easy to understand
- Easily implementable
- Can be implemented on different data standards [4].

Disadvantages:

- Ambiguity cannot be resolved
- Cannot deal with facts
- Mono-lingual technique

- Not easily adaptable on different platforms [4].

2.3.2 Pattern Learning Technique

This technique involves writing and editing patterns which requires a lot of skill. It also consumes a considerable amount of time. These patterns are not easily adaptable to new platforms used for different databases [4].

2.3.3 Supervised Learning Technique

This is pipeline style information extraction technique. In this method the task is split into different components and data annotation is prepared for these components. Several machine learning methods are used to address these components separately. Name tagging and relation extraction are some of the progress made using this field [4].

2.4 Web Data Extraction

Internet is a very powerful source of information. A lot many business applications depend on the internet for collecting information which plays a very crucial role in the decision making process. Using web data extraction we can analyze the current market trends, product details, price details etc. [7].

Web page generation is the process of combining data into a particular format. Web data extraction is the reverse process of web page generation. If multiple pages are given as input then the extraction target will be the page wide information and in case of a single page the extraction target will be record level information. Manual data extraction is time consuming and error prone [3].

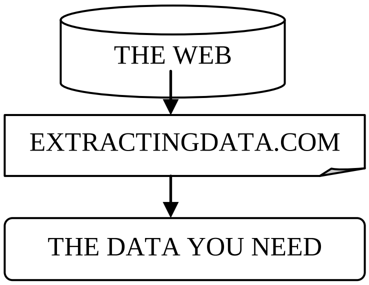

Figure: 2 web data extraction

2.4.1 WEB DATA EXTRACTION TOOLS

1) DELA (Data Extraction And Label Assignment For Web Databases)

DELA automatically extracts data from web site and assigns meaningful labels to data. This technique concentrates on pages that querying back end database using complex search forms other than using keywords [3].

DELA comprises of four basic components:-

a. A form crawler
b. Wrapper generator
c. Data aligner
d. Label assigner

❖ FORM CRAWLER

It collects labels of the website form elements. Most form elements contain text that helps users to understand the characteristics and semantics of the element. So, form elements are labeled by the descriptive text. These labels are compared with the attributes of the data extracted from the query-result page [3].

❖ WRAPPER GENERATOR

Pages gathered by the form crawler are given as input to the wrapper generator. Wrapper generator produces regular expression wrapper based on HTML tag structures of the page. If a page contains more than one instance of data objects then tags enclosing data objects may appear repeatedly. Wrapper generator considers each page as a sequence of tokens composed of HTML tags. Special token “text” is used to represent text string enclosed with in HTML tag pairs. Wrapper generator then extracts repeated HTML tag substring and introduces a regular expression wrapper according to some hierarchical relationship between them [3].

❖ DATA ALIGNER

Data aligner has two phases. They are data extraction and attribute separation [3].

➢ DATA EXTRACTION

This phase extracts data from web pages according to the wrapper produced by wrapper generator. Then it will load extracted data into a table. In data extraction phase we have regular expression pattern and token sequence that representing web page. A nondeterministic finite automation is constructed to match the occurrence of token sequences representing web pages. A data-tree will be constructed for each regular expression.

➢ ATTRIBUTE SEPARATION

Before attribute separation it is needed to remove all HTML tags. If several attributes are encoded in to one text string then they should be separated by special symbol(s) as separator. For instances "@", "$", "." are not valid separator. When several separators are found to be valid for one column, the attributes strings of this column are separated from beginning to end in the order of occurrence portion of each separator.

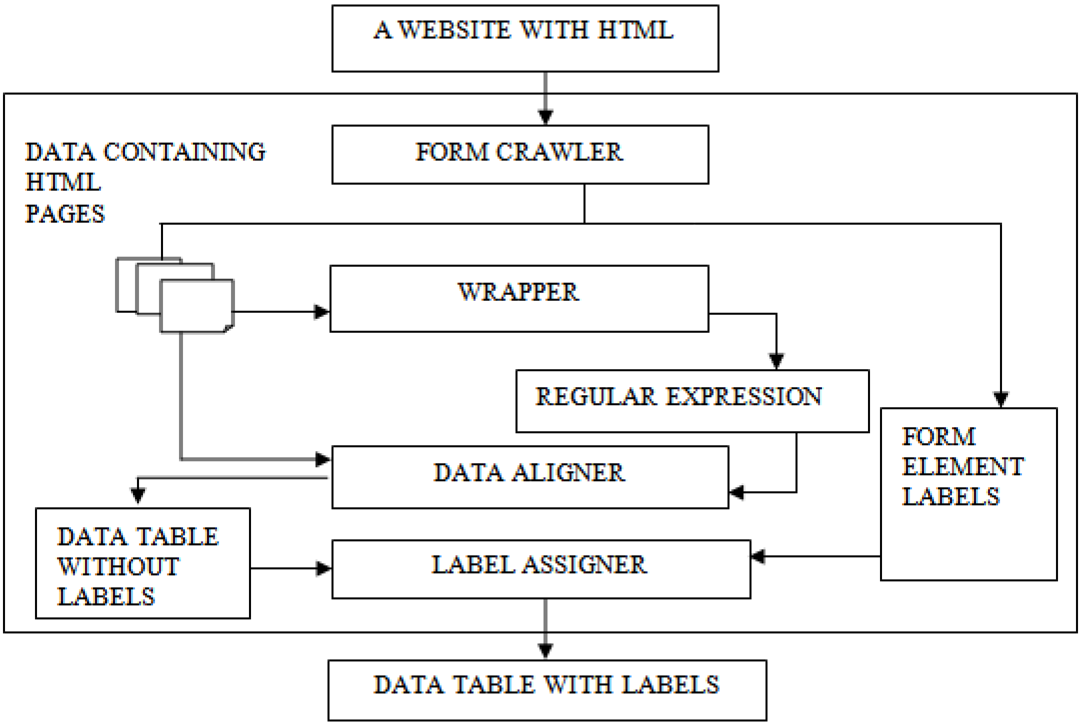

Figure: 3 DELA (Data extraction and label assignment for web data extraction)

2) FIFA TECH

Fifa Tech is a page level web data extraction technique. It comprises of two data modules through which data extraction is performed [3].

- First module takes DOM trees of web pages as input and merges all DOM trees into a structure called fixed/variant pattern tree.
- In the second module template and schema are detected from fixed/variant pattern tree.
- Peer node recognition: Peer nodes are identified and they are assigned same symbol.
- Matrix alignment: This step aligns peer matrix to produce a list of aligned nodes. Matrix alignment recognizes leaf nodes which represent data item.
- Optional node merging: This step recognizes optional nodes, the nodes which are which disappears in some column of the matrix.
- Schema detection: This module detects structure of the website i.e., identifying the schema and defining the template.

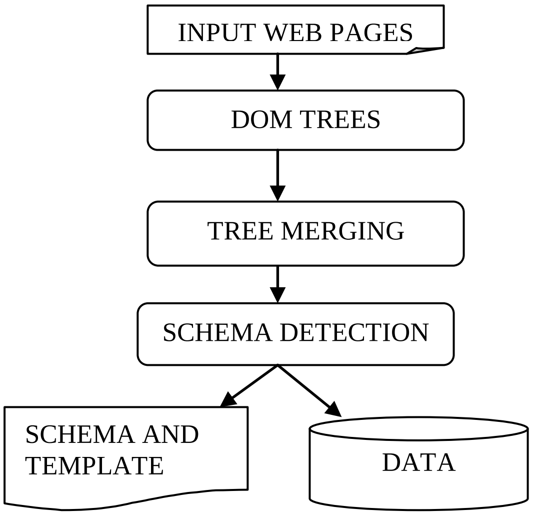

Figure: 4 Fifa tech for web data extraction

3) IEPAD

It is an information extraction system which applying pattern discovery techniques. It has three components, an extraction rule generator, pattern viewer and an extract module [3].

- Extraction rule generator accepts input web page and generate extraction rules. It includes a token translator, PAT tree constructor, pattern discoverer, a pattern validator and an extraction rule composer as shown in Fig.
- Pattern viewer is a graphical user interface which shows the repetitive pattern discovered.
- Extractor module extracts desired information from pages. [3]Translator generates tokens from input webpage. Each token is represented by a binary code of fixed length l.PAT tree constructor receives the binary file to construct a PAT tree. PAT tree is a PATRICIA tree (Practical Algorithm to Retrieve Information Coded in Alphanumeric). PAT tree is used for pattern discovery. Discoverer uses PAT tree to discover repetitive patterns called maximal repeats. Validator filters out undesired patterns from maximal repeats and generates candidate patterns.

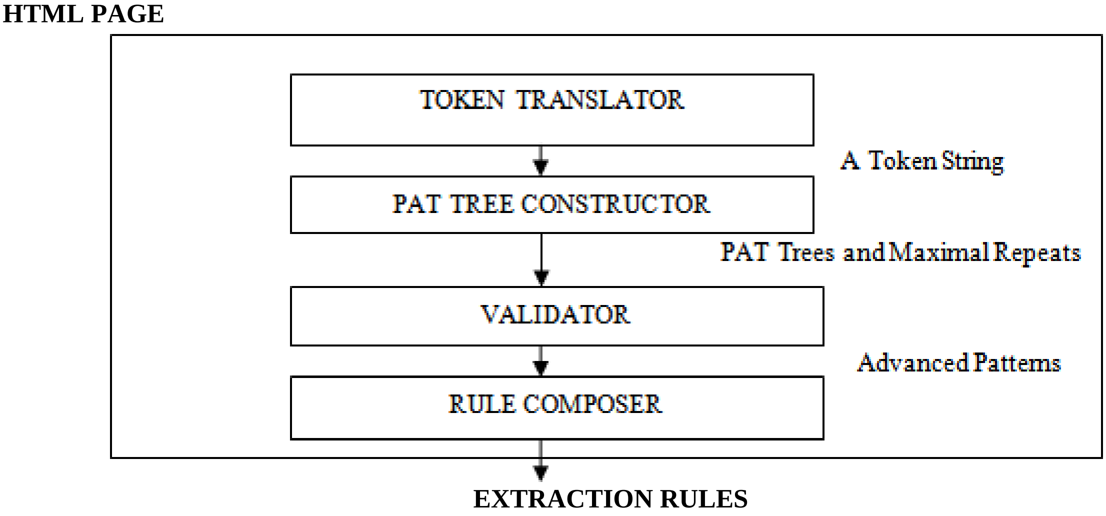

Figure: 5 IEPAD for web data extraction

III. Applications Of Information Extraction

Information extraction is used in different areas like business enterprises, personal applications, scientific applications etc. Information extraction plays a very important role in each and every field we work in. Various applications are listed as:-

3.1 Enterprise Applications

- **News Tracking**

  Information extraction plays a very important role in extracting information from different sources. There are two recent applications of information extraction on news articles: it integrates the data from videos and pictures of different events and entities of news articles and gathers background information on people, locations and companies [2].

- **Customer Care**

  A customer oriented enterprise requires to integrate itself with the customer requirements provided by them through mails and other means. This can be done by integrating customer interaction with the enterprise's own structured databases and business's ontologies [2].

- **Data Cleaning**

  Duplicacy of data needs to be removed from the databases. Data warehouse cleaning process is used to clean the information by keeping the similar data in same format at one place so that no redundancy arises. By dividing an entity into its properties it becomes easy to do the deduplication [2].

3.2 Personal Information Management

Personal information management systems aims to integrate the information provided in the form of e-mails, documents, projects and people in a structured format which links them with each other. Such systems are successful only if they are able to extract data from the predefined unstructured domains [2].

3.3 Scientific Applications

The rise in the field of bio-informatics has lead to the development of data extraction from terms like proteins and genes. Earlier it was not possible to extract the data from such biological terms but with the success in the field of information extraction and advancement in the techniques of data extraction it has become possible to extract data from various scientific entities and not stay limited only to the classical terms like people and companies [2].

3.4 Web Oriented Applications

- **Citation Databases**

  Many citation databases require different structure extraction steps for its creation. These are navigating websites for locating pages containing publication records, extracting individual publication records as per requirement of the database, extracting title, authors and references and segmenting all of them resulting in a structured database [2].

- **Opinion Databases**

  There is a lot of data available on the web related to any topic but in rough format. By using different structured techniques this data can be organized and all the reviews which lie behind the blogs, review sites, newspaper reports etc. can be extracted [2].

- **Community Websites**

  These are the websites which are created using data about researchers, conferences, projects and events which are related to a particular community. These structured databases require these steps for it's creation: locating the talks of the departments, finding the title of the conference, collecting the names of the speakers and so on [2].

- **Comparison Shopping**

  Different websites are created listing the information of their products and their price details. Such websites are used when looking for any product and its data is extracted from these collectively in order to form a structured database [2].

IV. Challenges and Future

- Traditionally the task of information extraction was carried out in only one language i.e., English. But with the growing textual data available in other languages there had become a need to extract data in all these languages [1].
- Designing information extraction techniques in different languages creates a difficulty in implementation. Hence, in order to remove this difficulty such techniques are designed which work for all languages [1].
- The different protocols to be used for information extraction are designed and formulated in such a way that all the textual data from different languages can be extracted easily [1].
- Thus, it is harder to implement information extraction using different languages and also the performance of non-English information extraction techniques is also lower [1].

- Extracting different entities from database has become easy by the different methodologies being developed but the identification of the relationship among these is still a very challenging task to perform. Bunescu and Mooney (2005b) have introduced a Statistical Relational Learning model to deal with this complex problem and are trying to investigate and find results on this side [5].
- In current researches, emphasis is also laid on the characteristic of extracting information not only from one single source but from multiple sources [5].
- Extracting facts from multiple sources helps in defining them and understanding them more precisely and accurately. Many efforts are being laid in this affect.
- As another future work, more emphasis is required to be laid on the applications of information extraction. Investigation on these applications will provide more sources of information extraction and can bring new challenges to this field. This is because different applications have different characteristics and different techniques are required to extract information from these fields.

V. Conclusion

This paper includes the study of information extraction systems in the past. Then comes the architecture of the knowledge extraction system and the different components included in it. It includes domain independent and domain specific components. Different knowledge extraction techniques are discussed like rule based technique, pattern learning technique and supervised learning technique. A method of data extraction, web data extraction, is detailed with its different tools. Tools like DELA (data extraction and label assignment), fifa tech and IEPAD have been discussed.

Extracting information from web plays a very important role in different business, personal and scientific applications. The next part of the paper includes the various applications of information extraction. The different techniques of information extraction helps in designing customer care applications, citation databases, news tracking applications etc. and in used for the purpose of data cleaning.

A vast progress has been made in this field but a more lot is still to be done. Researches continue in this field, emerging with new techniques with more efficiency and more performance. A model for information extraction system which acts as the base model for all the language dependent databases has been developed and removed the complex problem of developing a separate model for each database with different language as its base. A great future development is expected in this field with the researches continuing to move in more depth and bringing in new terms and techniques towards information extraction.

REFERENCES

[1] Jakub Piskorski and Yoman Yangarber (2013), *"Information extraction- past, present and future"*, The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Multisource, Multilingual Information Extraction and Summarization, Publisher: Springer, ISBN: 978-3-643-28568-4.

[2] Sunita Sarawagi (2008), *"Information extraction"*, Foundations and trends in databases, Volume.1, Issue No. 3(2007), 261-377, **DOI**: 10.1561/9781601981899, **E-ISBN**: 978-1-60198-189-9, **ISBN**: 978-1-60198-188-2.

[3] Devika K, Subu Surendran (April, 2013), *"An overview of web data extraction techniques"*, International journal of scientific engineering and technology, Volume 2, Issue 4, pp: 278-287, ISSN: 2277-1581.

[4] Heng ji (june 12,2012), *"Information extraction: techniques, advances and challenges"*, North American Chapter of the Association for Computational Linguistics (NAACL) Summer School.

[5] Jie Tang, Mingcai Hong, Duo Zhang, Bangyong Liang and Juanzi li (2007), *"Information extraction: Methodologies and applications"*, Emerging Technologies of Text Mining: Techniques and Applications, pp: 1-33.

[6] Douglas E. Appelt (1999), *"Introduction to information extraction"*, Artificial Intelligence centre, SRI International, Menlo Park, California, USA, ISSN: 0921-7126.

[7] Alexander Yates (2007), *"Information Extraction from the Web: Techniques and Applications"*, University of Washington.