Asian Journal of Information Technology 4 (12): 1187-1194, 2005

Automatic Ontology Generation for Semantic Search System Using Data Mining Techniques

K.R. Reshmy and S.K. Srivatsa
Sathyabama Institute of Science and Technology, Chennai, India

Abstract: Here we present about automatically generated ontologies for a semantic web search system using data mining techniques. This will improve the query process and will get better semantic results. Ranking algorithm is used to search and analyze web documents in a more flexible and effective way. Hyperlink structure of web document is utilized to rank the results. We use association rule mining to find the maximal keyword patterns. Clustering is used to group retrieved documents into distinct sets. This will extract knowledge about qury from the web,populate a knowledge base. The search engine that searches the web documents so far are syntactic oriented. Here we develop a searching system that semantically searches the documents. The semantics of the terms is achieved using the ontologies. Ontology serves as Meta data schemas, providing a controlled vocabulary of concepts, each with explicitly defined meaning. Ranking algorithm used here is the hyper textual ranking algorithm that scans both the contents of the documents and also the reciprocally linked documents. This technique has several advantages that include providing better semantic notion during the search. It also serves for multiple frame documents. There is a need for automatic generation of ontologies when using the semantic searching system. The paper here focuses on how the automatic generation of ontologies could be done for a semantic search system using datamining techniques.

Key words: Ontology, data mining, semantic web, meta data, web search, information retrieval, ontology population

INTRODUCTION

World wide web is the most excited society in the last 20 years. Web has turned to be the largest information source Available in the planet. It is the huge, explosive diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information and also raises the complexity of how to deal with the information from the different perspectives of view-users, web service providers, business analysts. The users want to have effective search tools to find relevant information easily and precisely. Web mining is the term of applying data mining techniques to automatically discover extract useful information from the web.

The web mining taxonomy: The web mining is has three fundamental dimensions, namely web content mining, web structure mining and web usage mining. The in-depth taxonomical classification is depicted in Fig. 1.

Web content mining: The web documents have heterogeneous structure which makes difficult to categorize, filter or interpret documents. So a more intelligent tool is needed for information retrieval such as intelligent agents. Moreover advanced database and data mining techniques are required to provide higher level organization of semi-structured data available on the web. Some efforts include.

Agent based approach: This involves development of sophisticated AI systems that can autonomously or semi-autonomously discover and organize web-based information on behalf of a particular user. The agent-based web mining system can be categorized as follows:

Intelligent search agent: This system has the characteristics of a particular domain to organize and interpret the discovered information. It relies on pre-defined and domain specific information about particular types of documents or on the coded models of the information sources to retrieve and interpret documents. Severalintelligent web agents have been developed that search for relevant information using domain characteristics and user profiles to organize and intepret the discovered information.Traditional search techniques use keywords as input to find the information that a user wants.But here only very little relevant portions only the user will get.here we use the datamining techniques to search the web documents in a more efficient way. We use a new hyper-textual ranking algorithm to look much deeper into the content of linked documents. If a ranking algorithm superficially considers

Corresponding Author: K.R. Reshmy, Sathyabama Institute of Science and Technology, Chennai, India

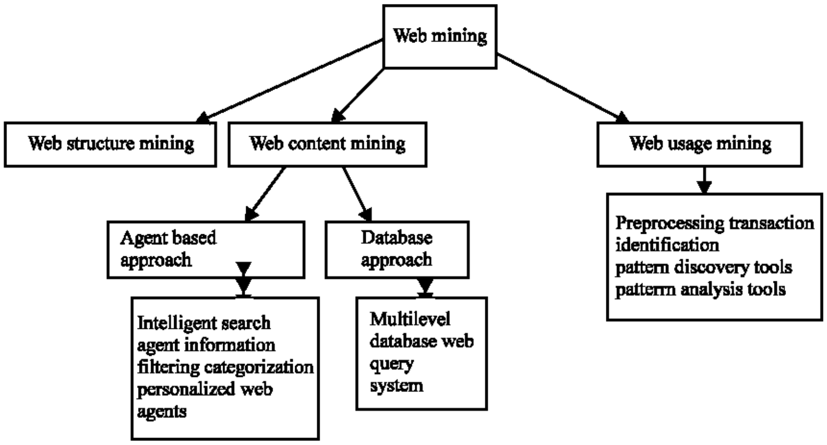
Web mining
Web structure mining
Web content mining
Web usage mining
Agent based approach
Database approach
Intelligent search agent information filtering categorization personalized web agents
Multilevel database web query system
Preprocessing transaction identification
pattern discovery tools
patterm analysis tools

Fig. 1: Web mining taxonomy

that all the linking documents fenerated automatically by web document publishing tools.To rank multiframe web documents also we can use the new ranking algorithm.The new ranking algorithm utilizes the hyperlink and hyper document information to address the need to include ranking criteria for rich and relevant content in the search for information on the web. Web mining is the data mining techniquesapplied on the web.we use three algorithms association rule mining, sequential pattern mining and clustering.Weighted association rule minng is used to retrieve the frequently accessed keywords in retrieved web documents.This is helpful since it also shows the relationships between data items. The best techniques to mine the online documents is clstering. That will segement the data into groups.so here we can use clustering to group related words together. Fuzzy C clustering algorithm is used to divide the retrieval documents into a user specified number of groups.

The technique we have used is the efficient web content mining. Here an intelligent searching algorithm based on hyper textual information is used. It is coupled with automatic generation of ontologies.

EXISTING TECHNOLOGIES

Every search engine must have 3 main components: crawler and indexer, searcher and ranker and interface.

Crawler and indexer: A crawler is also called as a robot, agent or spider.It is an unattended program that works continuously and automatically.It automatically scans the websites and collects web documents. From the links the crawlers find the other related documents The order to visit the linked documents is done by depth first and breadth first searches coupled with heuristics. Then the documents are indexed.

Searcher, ranker and interface: The user enters the keyword. Then the searcher scans the indexed documents matching the keyword.The ranker performs the ranking function and hence determines the order in which the document must be displayed to the user.It receives the queries and displays the result.The interface must be simple,intuitive and easy to use.

Ranking algorithms

Vector space model: Most of the engines use these ranking algorithms.It is based on the hyperlink structure (the quality measure of a web document is determined by counting the number of documents that has links to a document)

Page rank algorithm: This algorithm is used in google search Engine[1,2]

HITS: Hyper Link Induced Topic search

This algorithm finds the authoritative documents.these documents are said to be information rich. The algorithm also tracks the hub documents i.e. documents that have links to many authority documents. Some other ranking techniques include edge-weighted strategies, extended HITS etc.

Hyper textual ranking algorithm: The hyperlinks between the hyper documents contains useful information.This information is utilized in the hyper textual algorithm. The contents of the linked documents are also evaluated. It provides better semantic notion.

Reasons for selecting the algorithm

- It provides better semantic notion by retrieving relevant documents.
- It uses ontologies to identify and rank relevant web documented semantically.

- Thus the problems of polysemy, synonym and content sensitivity are prevented.
- It also ranks multi frame web documents.
- The text and hypertext are not just evaluated but also the contents in the reciprocally linked documents.

Data mining algorithms: To enhance user friendly searching some data mining algorithm are used. The 2 most significant algorithms are

Associate rule mining: These algorithm explorers the frequently used keywords set which can be used for subsequent querying.

Fuzzy c-Means clustering algorithm

Ranking process: The new Ranking Algorithm used here[1]. In this the compounded ranks into three categories of sorted order.These three categories are1) documents that contain all main search concepts[2].
documents that contain some of the main search concepts and have linkage relationships with other concepts[3] documents that have linkage relationships with some of the main search concepts.

Here hypertext characteristics of web documents and ontology are used to model the ranking algorithm to provide more flexibility.

PROPOSED ARCHITECTURE

The semantic web search system that uses the hyper textual ranking algorithm discussed so far has several advantages..The proposed intelligent search system has 6 main components:

They include crawler, language processor, interface, query engine miner and db connector. The crawler: It is also known as agent,an unattended program that works continuously,and automatically,having the essential role of locating information on the web and retrieving it for indexing.

Language processor: It is used by all theother components to process textual information.

Interface: It provides a user with a way to input query terms,request mining process and display query and mining results.

Query engine: It is the heart of the system.i t searches the inverted file indexes,which are created by the crawler in our index databasefor efficient retrieval of the documents matching the query terms provided by the user.It uses the linking structure of the retrieved documents to expand the query results.The new ranking algorithm[1] helps to display the order based on the degree of how well the results match the user's query.

Miner: It provides several kinds of data mining techniques.clustering groups the retrieved documents for a user's query.association rule mining uses the retrieved documents for more specific documents.

Data base connector: There are five databases that are connected by connection threads to enhance efficiency of the system. The databases are

- Index db
- Connectivity db
- Sentence db
- Oontology db
- Stop word db

Ontology: The ontology in this search system act as conceptual backbone for semantic document access by providing a common understanding and conceptualization of a domain.ontology consists of two main components; term and termrelationship. Term is the basic terms comprising the vocabulary of a domain.Term relationship is asset of relationship between terms.Populating ontologies with a high quantity and quality of instantiations is one of the main steps towards providing valuable and consistent ontology based knowledge services.Manual ontology population is very labour intesive and time consuming[4].Some semi automatic approache shave been presented.But are not adequate.Here we present a fully automatic approach of feeding the ontology with knowledge extracted from the web. Information is extracted with respect to a given ontology and provides XML files, one per document, using tags mapped directly from names of classes and relationships in that ontology.The foll Fig. shows an example of the XML representation of the extracted knowledge and how it is asserted in the ontology[5].

The system has been expanded to 8 components

Crawler: The crawler retrieves documents and sends for indexing to the index db. It has four modules retrieval, URL listing, formatting and modeling and hypertext parser.Retrieving module is sed to retrieve information from the web.This module fetches URLs from large storage of candidate URL's stored in the URL listing module.The hypertext parser module processes retrieved resourses. It will 1) determine the retrieved data type. 2) parsing the retrieved hypertext documents[6] extracting the hyperlinks and specified syructures in the documents.The

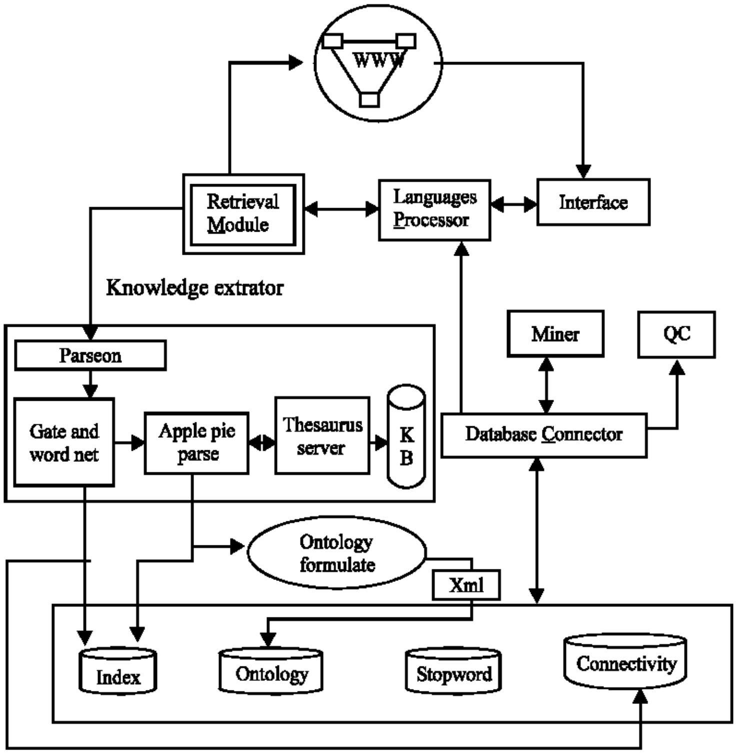

Fig. 2: Proposed system architecture

results are then passed to the Formatting and Indexing Module.The processing module adds the parsed URL'S to the URL listing module.The La nguage processor efficiently and effectively converts the retrieved text into uniform expression used for data mining and information retrieval.After converting the text the formatting and indexing module updates the index databaseto index the gathered web documents for later searches.,adds the hyperlink structure information to the connectivity data base and splits all the sentences in the acquired documents for the sentence data base.The URL listing module feeds thr retrieving module and makes some decisions for selecting URLs from the processing module to be added to the pol of candidate URLs.

Language processor: It processes the textual information for the other components. Our system can process data in English language.Three processes are there.Case translator,Word Stemmer and Stopword Filter To convert all the retrieved English words into a lower Case,case translater is used.Word Stemmer reduces words to their morphological root. Stopword filter removes insignificant words.

Knowledge extractor: Documents on the web use limiteless vocabularies,structures and composition styles. This make it hard for any IE technique to cover all variations of writing patterns.traditional IE systems lack the domain knowledge required to pick out relationships between the extracted entities.

Here ontology is coupled with a general purpose lexical database(word Net) and an entity recogniser (GATE) as guiding tools for identifying knowledge fragments consisting of not just entities,but also the relations between them.. Then performs knowledge extraction. The output of the extraction process is an XML representation of the facts, paragraphs, sentences and keywords identified in selected documents. It has

Parser: To parse documents into paragraphs and sentences.

Apple-pie parser: Groups grammatically related phrases to derive relationships.

GATE and word net: Identifies termsThesaurus server- to query the thesaurus knowledge base.

The following figure shows an example of knowledge extraction.

Db connector: It connects all the four data bases.

Index db: This db has term table and documents table. The documents table has document Id and posting id

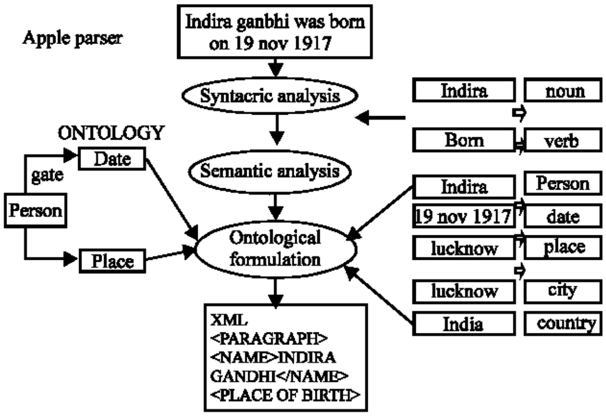

Fig. 3: Example for knowledge extraction

columns. Posting id represents the position of the document identified by the doc id. The term table has term id and posting id.Where the posting id specified the the doc id which has the corresponding term id.

Connectivity db: It also has two tables such that a doc-id can point to a given doc-id quickly and vice versa.

Stop word db: It has 615 insignificant terms for stop word filtering.

Ontology db: This db has domain specific ontologies. They are automatically fed with knowledge extracted from the web.

Thesaurus db: This has record of words and related words and their degree of semantic correlation.

Thesaurus server: To query the thesaurus knowledge base.

Query engine: The Query engine consists of ranker and searcher.it searches the inverted file indexes, which are created by the crawler in the index database for efficient retrieval of the documents matching the query terms provided by the user.It uses the linking structure of the retrieved documents to expand the query result. The ranking algorithm displays the order of web documents based on how well result matches the user query. The new Ranking Algorithm is used here[3]. In this the compounded ranks are separated into three categories of sorted order. These three categories are1) documents that contain all main search concepts[4] documents that contain some of the main search concepts and have linkage relationships with other concepts[3] documents that have linkage relationships with some of the main search concepts. Here hypertext characteristics of web documents and ontology are used to model the ranking algorithm to provide more flexibility.

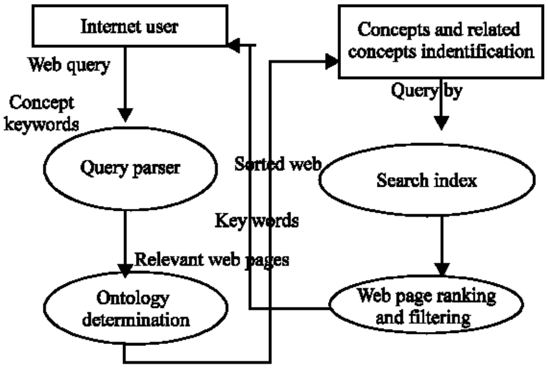

Fig. 4: Searching process

Miner: It provides several kinds of data mining techniques. Clustering groups the retrieved documents for a user's query.Association rule mining uses the retrieved documents for more specific documents.

The miner groups the retrieved documents in clusters as specified by the user. Then the top 10 representative keywords and corresponding documents are displayed. It also traces the frequently used keyword sets for subsequent usage.

SEARCHING PROCESS

The retrieval module retrieves URL s from URL listing module. The hyper text parser parses and fetches the links which are added to the listing pool. The gathered web documents are given to the index and forming module that will be send to the indexed databases for indexing. After a user inputs several keywords for searching relevant web documents, our searcher performs a lookup of the terms in the ontology databases to get the ontology containing these words. The Ontology will be created automatically. The searcher scans the search index in the index database for every key term in search concepts to obtain all the of the conceptually related documents. Then the ranker uses these documents and ontology for ranking and filtering on order to get a sorted document list for all of the relevant documents corresponding to the user's query.

ONTOLOGY

Ontology is the conceptual back bone for semantic document access Here ontology is created automatically.

Automatic ontology population: Manual ontology population is time consuming. Populating ontology with

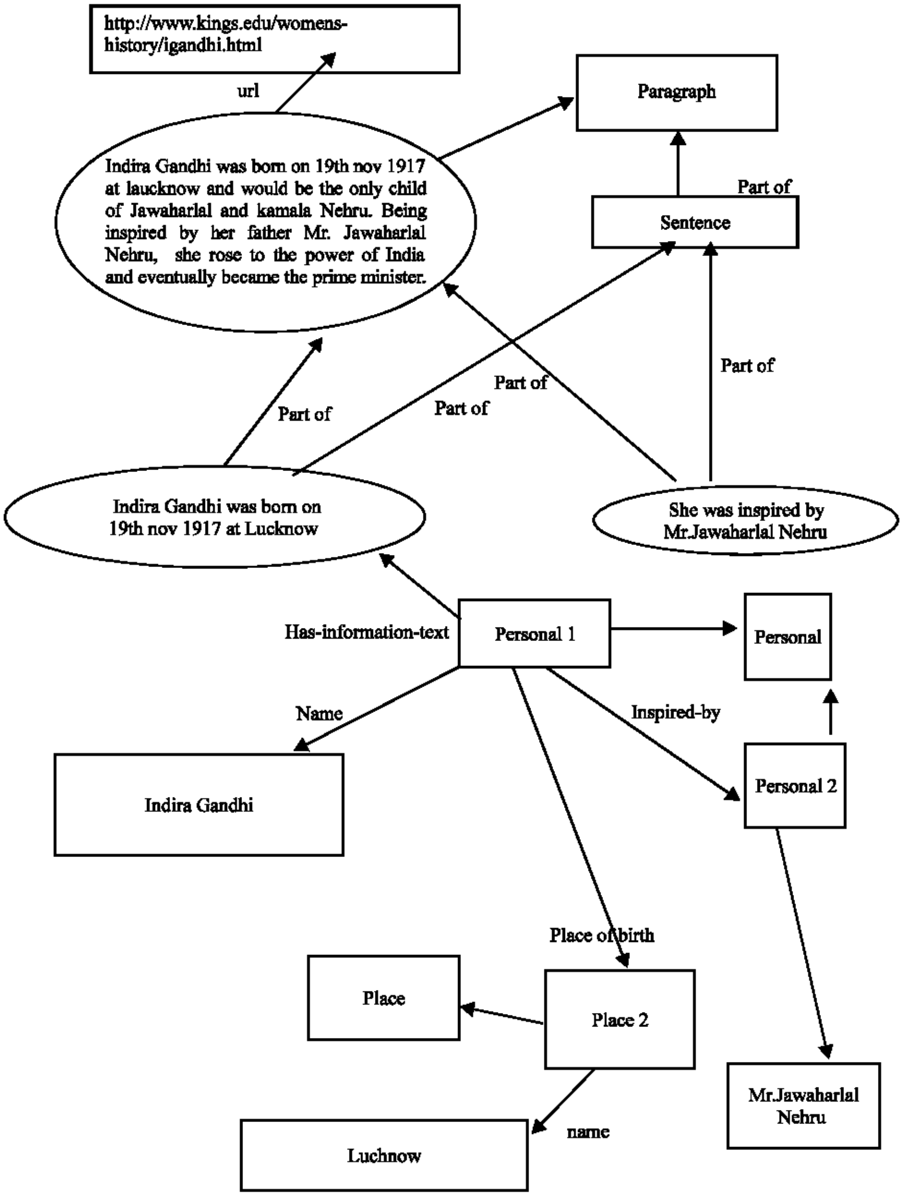
http://www.kings.edu/womens-history/igandhi.html
url
Paragraph
Indira Gandhi was born on 19th nov 1917 at laucknow and would be the only child of Jawaharlal and kamala Nehru. Being inspired by her father Mr. Jawaharlal Nehru, she rose to the power of India and eventually became the prime minister.
Part of
Sentence
Part of
Part of
Part of
Part of
Indira Gandhi was born on 19th nov 1917 at Lucknow
She was inspired by Mr.Jawaharlal Nehru
Has-information-text
Personal 1
Personal
Name
Inspired-by
Indira Gandhi
Personal 2
Place of birth
Place
Place 2
Mr.Jawaharlal Nehru
Luchnow
name

Fig. 5: Corresponding instances and relationships in the ontology

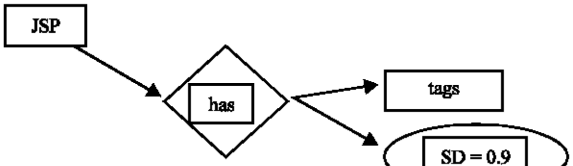

Fig. 6: Ontology representation for terms like JSP and tags

ahigh quantity and quality of instantiations is one of the main steps towards providing valuable and consistent ontology based knowledge services.The ontology is selected only if it already exists. If the term is new then human intervention is needed. So there is a need of automatic approach of feeding the ontology with knowledge extracted from the web. There are certain semi-automatic methods which are at infancy. In such cases document annotations were created and results were stored as assertions in an ontology.Here we present a fully automatic approach of feeding the ontology with knowledge extracted from the web. Information is extracted with respect to a given ontology and provides XML files, one per document, using tags mapped directly from names of classes and relationships in that ontology.

The foll.fig shows an example of the XML representation of the extracted knowledge and how it is asserted in the ontology.

```
<paragraph>
<url>http://www.kings.edu/womens-history/igandhi.html>
<text>Indira Gandhi was born on nov 19,1917 at Lucknow and would be the only child of Jawaharlal and Kamala Nehru.Being inspired by her father Mr.Jawaharlal Nehru,Ingira Gandhi rose to power in India and eventually became the Prime Mininster of INDIA.>
</text>
<sentence>
<text> Indira Gandhi was born on nov 19,1917 at Lucknow
</text>
<person>
<name>Indira Gandhi</name>
<place-of-birth>lucknow</ place-of-birth>
<date-of-birth>
<day>19</day>
<month>11</month>
<year>1917</year>
</ date-of-birth>
</person>
</sentence>
…..
<sentence>
<text>she was inspired by her father Mr.Jawaharlal Nehru</text>
<person>
<name>Indira Gandhi</name>
<inspired-by> Mr.Jawaharlal Nehru</inspired-by>
<person>
</sentence>
…….
</paragraph>
```

The URL, retrieved by the retrieval module is sent to the knowledge extractor. The extractor has a component called the parsor which extracts the sentences of each document. The GATE performs the extraction of the terms. Then the Apple Pie parser provides grouping of grammatically related phrases. Thus the relationships are obtains. The thesaurus db through thesaurus server will provide the semantic degree value. The thesaurus db has tables for terms, their synonyms and their semantic degrees. Thus the basic ontology has term, relationship and semantic degree.

The above diagram represents the basic ontology representation for the terms JSP and tags. The semantic degree is the value between -1 to 1 where it represents the relevancy of two terms (synonym) or irrelevancy (polysemy).Semantic degree is a property of the term ontology. The terms are updated into the index database. The term name, relationship form APPLEPIE and the SD from the thesaurus KB are updated into the ontology database. The terms with properties are formulated into an XML format and updated in the ontology database. Thus the new ontologies are generated automatically into the database to provide fast access to frequently used information via SQL queries.

CONCLUSION AND FUTURE WORK

The system we prodused is more effective and flexible for web serach with more semantic notion. We utilize hypertext characteristics of web documents and ontology to model the ranking algorithm to provide more flexibility.This system allows users to select the desired search domains that can correctly locate the documents they are looking for based on relevancy.The new ranking algorithm to look much deeper into the content of linked documents. In addition we used automatically generated ontology to solve traditional problems in text search that involve sysnonymy,polysemy and sensitivity. Data mining techniques are used to refine our search engine.Three useful techniques were used.association rule mining to explore primary keywords of retrieved documents,fuzzy c-means clustering to provide an overview of the desired documents.We implemented the system and tested it with english web documents from our university web sites.Preliminary results show that our web search system is effective and efficient.The system we prodused integrates a variety tools in order to automate an ontology – based knowledge acqusition process and maintain a KB.In the future we can improve the flexibility of our system even further.Issues of duplicate information across documents and reduntant annotations are still major challenges of automatic ontology population.Automatically populating an ontology from diverse and distributed web resources poses significant challenges.

REFERENCES

1. Chen, Yu.Ru., H. Ming-Chuan and Y. Don Lin, 2003. Using data mining to construct an intelligent web search system.
2. Millard, D.L. and K. Sanghe, 2003. Automatic knowledge based extraction and tailored biography generation from the web Harith Alani. IEEE Intelligent system, 18: 14-21.

3. Agrawal, R. and R. Srikant, 1995. Mining sequential patterns. In: Proc. 11th Intl. Conf. Data Engineering, pp: 3-14.
4. http://www.google.com/
5. http://www.yahoo.com/
6. Fciravegna, A.D., Y. Wilks and D. Petrelli, 2002. Timely and non-intrusive active document annotation via adaptive information extraction. Workshop on semantic authoring, Annotation and Knowledge Markup,15th European Conf. Artificial Intelligence Lyon, France, 2002.