A Frequent Pattern Based Approach to Information Retrieval

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Sanchika Gupta
(Roll No. 800932019)

Under the supervision of:
Dr. Deepak Garg
Assistant Professor (CSED)

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

June 2011

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "**A Frequent Pattern Based Approach to Information Retrieval"**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Deepak Garg* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Signature:
(**Sanchika Gupta**)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(**Dr. Deepak Garg**)
Computer Science and Engineering Department,
Thapar University,
Patiala

Countersigned by

(Dr. Maninder Singh)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

ACKNOWLEDGEMENT

It is a great pleasure for me to acknowledge the guidance, assistance and help I have received from Dr. Deepak Garg. I am thankful for his continual support, encouragement, and invaluable suggestions. He not only provided me help whenever needed, but also the resources required to complete this thesis report on time.

I am also thankful to Dr. Maninder Singh for his kind help and cooperation.

I would also like to thank all the staff members of Computer Science and Engineering Department for providing me all the facilities required for the completion of my thesis work.

I would like to say thanks for support of my classmates. I want to express my appreciation to every person who contributed with either inspirational or actual work to this thesis.

I am highly grateful to my parents and brother for the inspiration and ever encouraging moral support, which enabled me to pursue my studies.

Sanchika Gupta

ABSTRACT

The intent of this thesis is to discuss the information retrieval system for local databases. The approach is to search the web both semantically and syntactically. The proposal handles the search queries related to the user who is interested in the focused results regarding a product with some specific characteristics. The objective of the work will be to find and retrieve the accurate information from the available information warehouse which contains related data having common keywords. This information retrieval system can eventually be used for accessing the internet also. So both semantic and syntactic search engine are compared for information retrieval using two parameters i.e. precision and recall.

First of all a comparison is made for which both semantic and syntactic information retrieval system are created. For the syntactic information retrieval system, a local database is created which consists of related information and a simple search engine is developed to retrieve information from that database. For the semantic information retrieval system, ontology with same information as in the database is created and queries are used to extract information from that. Accuracy in information retrieval that is achieving both high precision and recall is difficult. Different parameters used to measure the effectiveness of the information retrieval system include precision, recall and various fundamental search facilities like symbols, keywords, wild cards, case sensitiveness etc. The final analysis has been depicted via comparison table and line graph.

After that based on the results, an ontology based Information retrieval for learning styles of autistic people is created. This proposal handles the search queries related to the autistic people based on their learning styles, so that user may get focused results and unwanted/irrelevant data can be minimized.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

1. IR: Information Retrieval
2. OS: Operating System
3. FTP: File Transfer Protocol
4. GUI: Graphical User Interface
5. OWL: Ontology Web Language
6. DL: Data Logic
7. IBM: International Business Machines
8. W3C: World Wide Web Consortium
9. AAC: Augmentative and Alternative Communication
10. YAACK: AAC *(*Augmentative and Alternative Communication*)* Connecting Young Kids
11. VCR: Video Cassette Recording
12. RDF: Resource Description Framework
13. WAMP: Windows, Apache, mySQL, PHP
14. PHP: Personal Home Page or Hypertext Processor
15. XP: eXPerience in Windows XP
16. CO-ODE: Collaborative Open Ontology Development Environment
17. ASLIB: The Association for Information Management

Chapter- 1
INTRODUCTION

1.1 Information Retrieval

IR is the area of study concerned with searching for documents, for information within documents, and for metadata about documents and finding material of an unstructured nature that is used to satisfy an information need from large data storage [1]. IR is also used to facilitate semi structured search, for e.g. finding such a document where the title contains oracle and the body contains threading and clustering which is the task of grouping the documents based on their contents. Before the retrieval process can even be initiated, it is necessary to define the text database which consists of related information from which information is to be retrieved. An IR process begins when a user supplies query to the system. Queries can be expanded as formal statements that defines the information needs, for example search strings that are used in web search engines. In IR, it is not necessary that in a collection of objects a query may identify a single object. Instead, several objects may match the query, probably with different degrees of relevancy. A query is a form of questioning i.e. in the form of inquiry. The style and format of querying might be different for both syntactic and semantic search engine.

1.2 Information Retrieval System

IR systems can be differentiated by the level at which these systems operate, and it is important to make a comparison between three basic levels. When a search is made over the web, the search engine provides search over zillions of documents that are stored on millions of hard disks. First of all is the personal IR. In the last many years, consumer operating systems have been integrated for e.g. Apple has provided us with Mac OS X Spotlight and Windows has provided Vista Instant Search. Moreover e-mail programs are very common these days. They along with providing searching facility also provide facility for text classification which is used to differentiate between spams, junk mails, work related mails and other personnel mails which can then be placed in particular folders. After the personal IR is the enterprise, institutional, and domain-specific search. For the enterprise search, the retrieval is provided for corporation's internal documents and for institutional search, a database of patents may be retrieved. Domain specific search can provide collections of research articles on a particular topic for e.g. on computer science. In these cases, the documents will be stored on centralized file systems rather than distributed and a large number of machines will provide search over the collection [1].

1.2.1 Syntactic Information Retrieval System

Every language has its own Syntax and Semantics. Syntax is the study of grammar. Semantics is the study of meaning. Syntax is how to say something. Semantic is the meaning behind what you say. Different syntaxes may have the same semantic: x += y, x=x+y. Syntax and semantics are all about communication.

A web search engine or the syntactic information retrieval system is designed to search for information on the WWW and FTP servers. The search results are presented in a list of results and are called hits. The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input.

1.2.2 Semantic Information Retrieval System

A semantic information retrieval system attempts to make sense of search results based on context. It automatically identifies the concepts structuring the texts. For instance, if you search for "passport" a semantic information retrieval system might retrieve documents containing the words "visa", "embassy" and "flights". Semantic web help computers understand and interpret information and also finds additional information that might be useful. What this means is that the search engine through natural language processing will know whether you are looking for a small animal or a Chinese zodiac sign when you search for “rabbit”.

1.3 Ontology

Ontology is a shared and common understanding of some domain that can be communicated between people and application systems. Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner. It defines the vocabulary with which queries and assertions are exchanged among domains. Ontologies have been widely accepted as the primary method of representing knowledge in the Semantic Web [2].

Ontology can be defined as formal and explicit specification of a shared conceptualization. In the above statement, conceptualization refers to an abstract model of phenomena that have been identified by the relevant concepts of those phenomena. Explicit means the type of concepts which are used, and the constraints on their use. Formal is referred to the ontology that is to be machine readable. Shared reflects the ontology that should capture consensual knowledge which is accepted by the communities. Ontology is the most important factor which enables interoperability in the Semantic Web [3]. The main problem with the use of a shared vocabulary according to a specific conceptualization of the world is that much of the information remains implicit. Ontologies have been set out to overcome the problem of implicit and hidden knowledge by making the conceptualization of a domain explicit.

Ontologies also facilitate machine processing by allowing information to be annotated with metadata so that meaning can be determined. Ontologies are expected to play a central role in the development of the Semantic Web, and will be used for many different purposes, such as querying, presentation, and navigation. Ontologies also include information about the relationships among terms (concepts), which provides the basis for semantic reasoning.

1.4 Semantic Web and Ontology

The Semantic web proposes to help computers understand and use the web. Metadata is added to web pages that can make the existing syntactic web machine readable. The main purpose of the Semantic web is driving the evolution of the current web by allowing users to use it to its full potential, thus allowing them to find, share, and combine information more easily. This won't bestow artificial intelligence or make computers self-aware, but it will give machines tools to find, exchange and, to a limited extent, interpret information.

The Semantic web combined with ontology can be used for visualization techniques in several different ways, but the visualization is dependent on characteristics of the ontology used. Ontology helps both people and machines communicate more effectively by providing a common definition of a domain. The GUI serves as an interface between the user and the system. OWL is the language used for developing ontologies. Figure 1.1 shows the ontology based semantic information retrieval system.

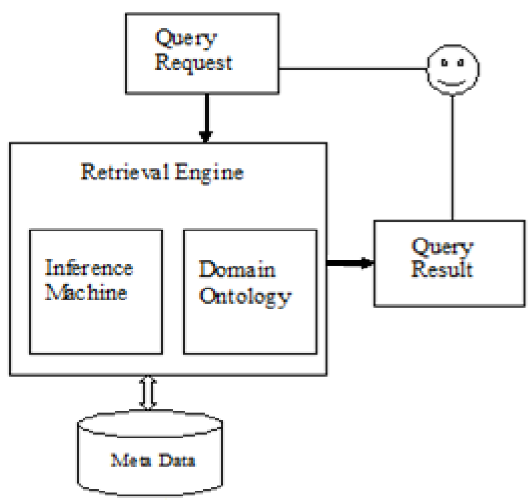

Figure 1.1: Semantic Information Retrieval using Ontology.

1.4.1 OWL Properties

OWL Properties represent relationships. There are three types of properties-

- Object properties- Object properties depicts the relationships between two individuals.

Figure 1.2: An object property linking the individual A to individual B.

- Datatype properties

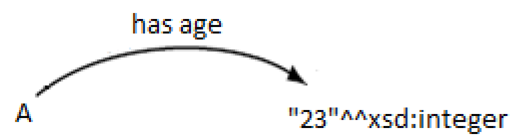

Figure 1.3: A datatype property linking the individual A to data literal ‘23’, which is a type of integer.

- Annotation properties- Annotation properties can be used to add information (metadata — data about data) to classes, individuals and object/datatype properties [4].

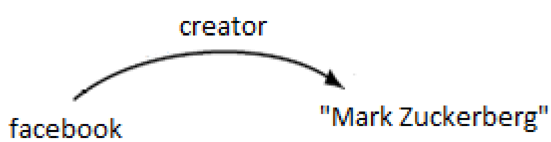

Figure 1.4: An annotation property, linking the class ‘facebook’ to the data literal (string) “Mark Zuckerberg”.

1.4.2 Restrictions

Restrictions now use infix syntax rather than prefix syntax. The description logic symbols that were previously used to indicate the type of restriction have been replaced with English language keywords.

Table 1.1: Description logic symbols and the corresponding English language keywords [4].

OWL	**DL Symbol**	**Manchester OWL Syntax Keyword**	**Example**
someValuesFrom	$\exists$	some	hasChild some Man
allValuesFrom	$\forall$	only	hasSibling only man
hasValue	$\ni$	value	hasCountryOfOrigin value England
minCardinality	$\geq$	min	hasChild min 3
Cardinality	$=$	exactly	hasChild exactly 3
maxCardinality	$\leq$	max	hasChild max 3

1.4.3 Data values and Datatypes (OWL1.1)

Constants can be expressed without type by just enclosing them in double quotes, or with type hasAge **value** "21"^^long

Usage of these datatypes in more general expressions is possible through their shortened name hasAge **some** int Person **and** hasAge **some** int[>= 18, <= 30].

Table 1.2: Data values and Data types [4].

XSD facet	Meaning
< x, <= x	less than, less than or equal to x
> x, >= x	greater than, greater than or equal to x
length x	For strings, the number of characters must be equal to x

maxLength x	For strings, the number of characters must be less than or equal to x
minLength x	For strings, the number of characters must be greater than or equal to x
pattern regexp	The lexical representation of the value must match the regular expression, regexp
totalDigits x	Number can be expressed in x characters
fractionDigits x	Part of the number to the right of the decimal place can be expressed in x characters

1.5 Autism

Autism is a lifelong developmental disability that first appears during infancy or childhood, and generally follows a steady course without remission. It affects how a person communicates with, and relates to, other people. It also affects how they make sense of the world around them. Overt symptoms gradually begin after the age of six months, become established by age two or three years, and tend to continue through adulthood, although often in more muted form. Autism interferes with the normal development of the brain. The three main areas of difficulty which all people with autism share are sometimes known as the 'triad of impairments'. They are:

- difficulty with social communication
- difficulty with social interaction
- difficulty with social imagination.

Difficulties concerning social communication include the following:

- Immature language skills.

- Communication with gestures.
- Not able to understand the facial expressions or tone of voice.
- Inability to attach meaning to words.

As for difficulty with social interaction, individuals with autism:

- Prefer to spend time alone rather than with others.
- Doesn’t understand the unwritten social rules.
- Are less comfortable with other people so show little interest in making friends.
- Are less responsible to social gestures such as eye contact.
- Are insensitive to others felling and anxious.

As for difficulty with social imagination is concerned it’s hard for autistic people to:

- Comprehend abstract ideas.
- Understand the concept of danger.
- Solve problems on their own.
- Distinguish between right and wrong.
- Imagine and plan for future.

1.6 Learning Styles of Autistic People

‘Learning styles' is a concept which attempts to describe the methods by which people gain information about their environment and that help someone learn and retain a skill or fact. Some individuals believe that spatial methods (such as social stories for learning social skills) are the only way to go with autistic children. But some autistic children have difficulty in understanding a pictorial concept. Some people believe that the only way to deliver skills training is by utilizing words, when perhaps for that child a repertoire of pictures may produce the best results. The teachers, diagnosticians and therapists are investing the time to find the most useful learning style for a challenged learner in order to put classroom, testing or therapy time to best use. People can learn through seeing (visually), hearing (auditory), and/or through touching or manipulating an object (kinaesthetically or 'hands-on' learning). For example, viewing a book of pictures or reading a textbook involves learning through vision, listening to a lecture in a classroom or on tape involves learning through hearing, and pressing buttons to determine how to operate a computer involves hands-on learning [5]. The predominantly used learning styles are:

- Visual (spatial) - using pictures, images, and spatial understanding.
- Aural (auditory-musical) - using sound and music.
- Verbal (linguistic) - using words, both in speech and writing.
- Physical (kinaesthetic or tactical) - using your body, hands and sense of touch.
- Logical (mathematical) - using logic, reasoning and systems.
- Social (interpersonal) - learning in groups or with other people.
- Solitary (by self) - working alone and use self-study.

Chapter- 2
LITERATURE REVIEW

Personal computers that have advanced by leaps and bounds, it has become cheap enough to be ubiquitous in the last thirty years. Moreover assistive computer hardware and software tools have served a great deal to bring to students with disabilities a whole new world of learning opportunities.

2.1 Assistive Technologies

Autistic children are motivated by imagination and consistent growth that is why computers can be thought of as the perfect assistive technology for learning. They make the child learn to function independently by putting the child in the driver's seat. It has been proved in a research that autistic children who use computers have increased attention spans, can stay longer in their seats, have developed fine motor skills and can repeat a behaviour that was learned at school that depicts a greater ability to generalize skills across environments. From the above it is inferred that computers has provided greater benefits, so they should be an integral part of a special education student's daily curriculum.

Moreover computer games can help children with autism in educational areas such as learning new vocabulary, practicing math skills or improving eye-hand coordination. Educational computer games that are developed for autistic and non-autistic children are appropriate. These games provide many educational benefits for autistic children. There are a variety of single-player and multiple-player games that can improve important developmental and behavioural skills.

One of the most prevalent technologies for developing user friendly and efficient tools for assistance is the Graphical User Interface (GUI), which was developed by Douglas Engelbart in 1960's. Moreover it is said that while the GUI was not designed for people with disabilities, it made the computer more accessible for them just as it has made the computer more accessible to the general population [6].

After the GUI was developed, IBM developed various technologies for the assistance of disabled people. The following is a list of tools developed by the company: Braille Printer, a talking typewriter for blind people, a talking display terminal, a screen reader for the sight impaired, a speech recognizer (Via Voice) that enables users to interact with the computer using voice commands and a talking web browser (Home Page Reader) for the visually-impaired and the elderly.

Various versions of Microsoft's Windows OS, in particular Windows XP, have included a range of accessibility options which are designed specifically to assists users with vision, hearing, and mobility impairments [7].

After the invention of WWW in 1991, the World Wide Web Consortium (W3C) has defined accessibility guidelines for the World Wide Web which requires the Web content provider to be aware of assistive technologies designed to facilitate access to the Web for users who have a disability [8].

Besides, there are various other companies and institutions that supply assistive products for people with varying disabilities. Among these are a group of products that are referred to as Augmentative and Alternative Communication (AAC) tools [9].

2.2 Background Knowledge about Autism

Autism is a developmental disability that typically appears during the first three years of the life. Being the result of a neurological disorder that affects functioning of the brain, autism and its associated behaviours occur approximately in 15 of every 10,000 individuals. Autism interferes with the normal development of the brain in communicative, cognitive and social areas. Below are the deficiencies faced by autistic people:

Problems concerning communication include the following:

• under-developed language skills;

• use of words without attaching the usual meaning to them;

• communication with gestures instead of words; and

• short attention spans.

Following is the list of cognitive problems exhibited by individuals with autism [10]:

• poor comprehension of abstract concepts;

• too-literal interpretation of information;

• poor comprehension of figures of speech;

• diminished ability to solve problems; and

• inability to distinguish pertinent information and stimuli from the irrelevant.

Many problems that autistic children experience can be alleviated through special education. The main goals when treating children with autism are to lessen associated difficulties and to increase quality of life and functional independence. No single treatment is adequate and treatment is typically tailored to the child's unique needs. Families and the educational system are the main resources for treatment [11].

2.3 Augmentative Communication

One way to improve the lives of autistic people is to introduce learning tools called Augmentative Communication [12]. This technology provides visual information in a creative format which helps a child in completing activities of daily living both in the home and at school. For example, an Activity Schedule can be created to remind the child what they have to do when they wake up to get ready for school. The morning routine can be represented using various photographs, drawings or words placed sequentially. E.g.

Page 1: a toothbrush and toothpaste;

Page 2: a washroom;

Page 3: a hairbrush;

Page 4: a shirt, sneakers, socks, and shoes;

Page 5: a bowl of fruits and milk;

Page 6: a backpack, lunch box, bottle.

After completing each of the above actions it is very important for the child to cross off the item or place the picture in a "completed" envelope. It is very helpful for the autistic child to understand the sequence of day to day events, when the information is presented in this manner.

As can be inferred from the work of many researchers (such as [3, 13, 14]), individuals with autism think in pictures, not words, and play a video in their mind when reasoning. Below is the pictorial representation of phases of an eating event.

Figure 2.1: Phases of an eating event.

One of the most remarkable mysteries of autism has been the magical ability of most autistic people to perform outstandingly at visual spatial skills while performing so poorly at verbal skills. The system is meant to assist an autistic person and motivate him when carrying out a reasoning process. Visual thinking enables them to build entire systems in their imagination. They prefer to use diagrams, pictures, study notes, handouts and movies to see the information they are learning. Temple Grandin, an architect with autism, often says that she think in pictures. She can translate both spoken and written words into full colour movies, complete with sound, which run like a VCR tape in her mind, which is difficult to understand for language based thinkers. This phenomenon helps her in her job as an equipment designer for the livestock industry. Visual thinking is a tremendous advantage for autistic people. That is, when somebody communicates with the autistic people, their words automatically get translated into pictures [11].

2.4 Crawler based Search Engine

The term ”search engine” is used to indicate both crawler based search engines and manually maintained directories, although they gather their indexes in radically different ways.

Here we are discussing the Crawler-based search engines, which are based upon the syntactic information retrieval system, such as Google which create their catalogues automatically: they crawl the web, then the user searches through what they have found. On the contrary, a manually maintained directory, such as the Open Directory, depends on humans: people submits a short description to the system about a certain site, or appropriate editors write a review for their assigned sites; thus, a web search looks for matches only in the submitted descriptions [15].

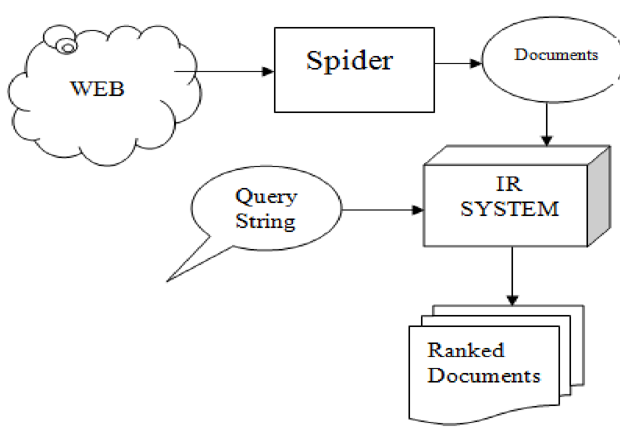

Figure 2.2: Syntactic Information Retrieval System.

Figure 4.4 depicts the syntactic information retrieval system. It consists of a spider which is a computer program that browses the web in an orderly fashion. It automatically discovers and collects resources, especially the web pages, from the Internet. This process is called spidering. Many search engine use spidering to provide up to date data. It provides a copy of all the documents which has already been visited for faster searches. So when user inputs a query string in the syntactic information retrieval system, the system then provides a list of ranked documents, ordered according to the requirement of the user to get high precision and recall.

Present day research on semantic web is concentrated on two main things: RDF and ontology. Ontology is a theory about the nature of existence, of what types of things exist. It is a study of such research. Ontology is a document or file that formally defines the relations among terms. The most typical kind of ontology for the web has taxonomy and a set of inference rules. The taxonomy defines classes of objects and relations among them.

Chapter- 3
PROBLEM STATEMENT

3.1 Problem Definition

Accuracy in information retrieval that is achieving both high precision and recall is difficult. A technique for information retrieval needs to be developed in which retrieved data is precise i.e. an IR system which handles the search queries related to the user who is interested in the focused results regarding a product with some specific characteristics.

A product may have different characteristics like price, weight, size, length, color, functionality based parameters etc. User may type the search queries using some words or a paragraph and should get focused results in which unwanted/irrelevant data is minimized. The objective of the work will be to find and retrieve the accurate information from the available information warehouse which contains related data having common keywords. This information retrieval system can eventually be used for accessing the internet also.

These days the information retrieval system which is used consists of a large amount of irrelevant and unwanted data, which needs to be minimized. So for that different techniques for information retrieval need to be studied and a comparison need to be made between them on the basis of different parameters, so that relevancy and accuracy of the retrieved data can be justified. Moreover a technique needs to be found which is simpler to use and consists of fundamental search facilities, so that it is easy for the user to search for the required item.

3.2 Gap Analysis

- The presently available information retrieval system consists of large amount of unwanted and irrelevant data.
- Accuracy in IR is difficult to achieve.
- The search engines present does not provide focused results for a given query.
- Relevancy of the retrieved data sets cannot be justified.
- Achieving high precision i.e. no of relevant items to the total number of items retrieved is not easy.
- At the same time, high recall ratio i.e. no of relevant items retrieved to the total number of relevant items in the database is also difficult to achieve.
- Different search engines using various techniques for retrieval of the data from the database does not consists of all fundamental search facilities.

3.3 Proposed Objectives

The problems or the limitations defined in the above section of this chapter are proposed to be solved by:

1. To observe and study the effect of various existing IR techniques.
2. To make a comparison between various IR techniques analyzed above.
3. To collect the datasets for verifying the accuracy and relevancy.
4. To find the results of the comparison by applying various IR techniques studied above on the datasets collected and identifying them on the basis of different parameters.

5. On the basis of the comparison results, to design a new information retrieval technique, that is able to solve the above problems.

3.4 Methodology

- To study the semantic and syntactic methods for information retrieval.
- To study ontology for semantic information retrieval.
- To work on an ontology tool using OWL language to develop a vocabulary for the search engine.
- To develop a search engine using WAMP server, PHP and MySQL for a local database.
- To design the semantic information retrieval system and analyze the syntactic information retrieval system.
- To compare the results of semantic and syntactic search for information retrieval using the same dataset for both the systems.
  The parameters used for comparison are:
  - Fundamental Search Facilities
  - Precision
  - Recall
- On the basis of the comparison results, to design an information retrieval system for the autistic people, that is easy to use and cater to the demands of these people.

3.5 Importance

Accuracy in information retrieval is a very crucial task to find the relevancy of the retrieved datasets. It is very important that the users of the web are satisfied with the search results. If a user enters a search query and the results presented are hardly helpful then it results in wastage of user's time and resources. So if such a technique exists which has a high accuracy rate then users are greatly benefited. It will also result in increasing the precision and recall ratio which results in minimizing the unwanted and irrelevant data. It is very common these days that a single term has multiple meanings, for e.g. “jaguar” which is a car, an operating system and an animal also. So a technique needs to be found that simplifies the searching and understands the user’s needs. Because till now no single technique is able to overcome the limitation, so this comparison will help to make an estimation that which technique is better or what particular properties of a technique are more efficient.

Chapter- 4
IMPLEMENTATION

First of all both the semantic and syntactic information retrieval system are created. Accuracy in information retrieval that is achieving both high precision and recall is difficult. So both semantic and syntactic search engine are compared for information retrieval using two parameters i.e. precision and recall.

For the syntactic information retrieval system, a local database is created which consists of related information and a simple search engine is developed to retrieve information from that database. For the semantic information retrieval system, ontology with same information as in the database is created and queries are used to extract information from that.

4.1 Semantic Information Retrieval System

The "Mediawiki Ontology" consists of information which is related. It helps to find information that is being searched and also provides the related information that might be helpful. Imagine this scenario. You want to purchase a car. You have heard about “jaguar” and want to know more about it, so you search for the term using your favorite search engine. Unfortunately, the results you’re presented with are hardly helpful. There are listings for jaguar the animal, a cat species etc. Only after sifting through multiple listings and reading through the linked pages are you able to find information about the Tata Group production “Jaguar”. On the other hand, the semantic information retrieval system can interpret and understand what is being searched for. The semantic web agent helps you to find the required car and also tells you about its features, functions, price and other available options.

Tool used to develop the ontology for semantic information retrieval system is Protégé. It is open source available on Website of Stanford University. Figure 4.1 shows the snapshot of class view annotation for mediawiki ontology.

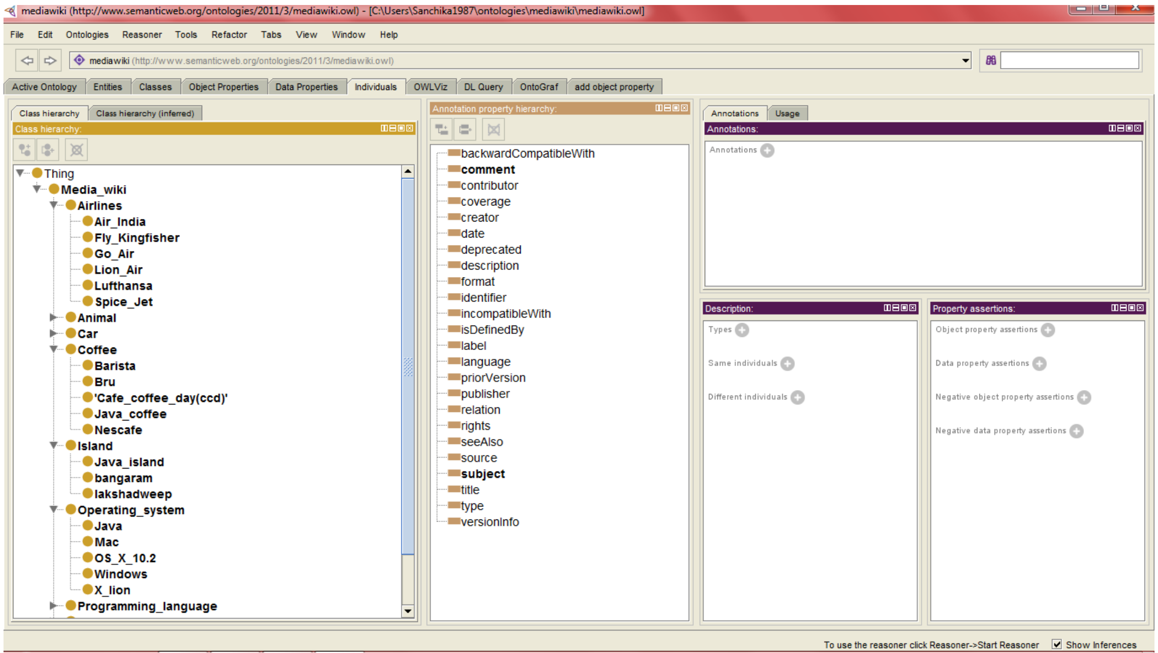

Figure 4.1: Class view Annotation

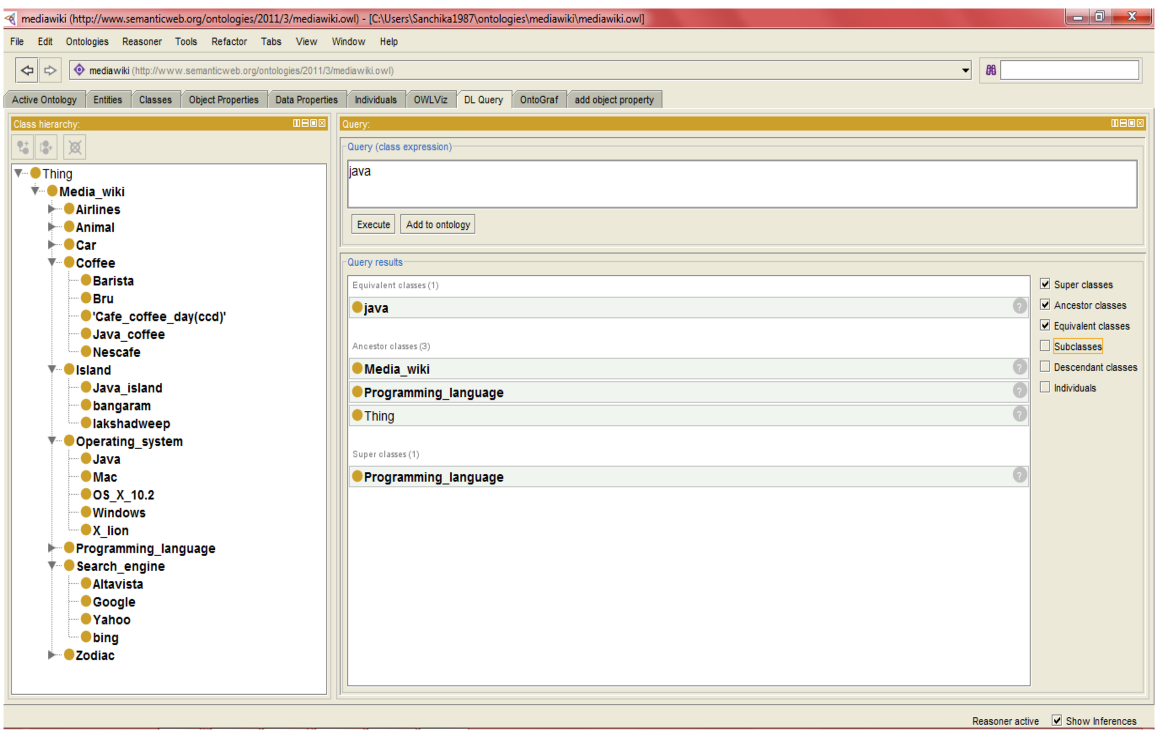

Figure 4.2: Snapshot of DL query showing the Super class, Ancestor class and Equivalent class of class expression “java”.

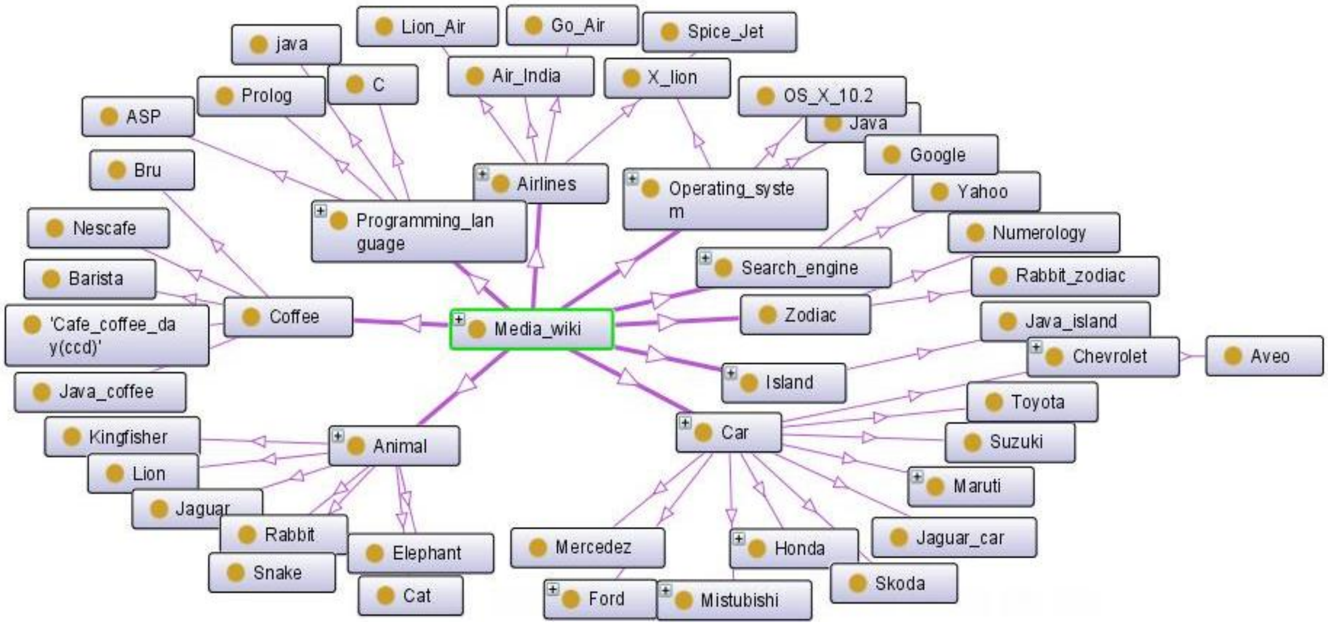

Figure 4.3: Mediawiki Ontology graph

Above Figure 4.3 presents the media wiki ontology graph which consists of related information.

4.2 Syntactic Information Retrieval System

In the syntactic information retrieval system a database is created "mediawiki" which consists of same information that is used to build the ontology, and a search engine is implemented using PHP to search from that database. This search engine retrieves every occurrence of the search item from the database [17]. For e.g. If we searched for "Lion", then every occurrence of Lion i.e. "Lion- the panther", "X Lion- Apple Mac operating system" and "Lion Air- Indonesia's largest private carrier airplane" is retrieved. Along with the search items links are also provided to get more information about them from the internet. This is implemented using the WAMP server. Figure 4.4 gives a snapshot of MySQL database in which all the related data is stored, Figure 4.5 gives a snapshot of the search engine page where the query has to be entered and Figure 4.6 gives a snapshot of search retrieval page when searched for "jaguar car".

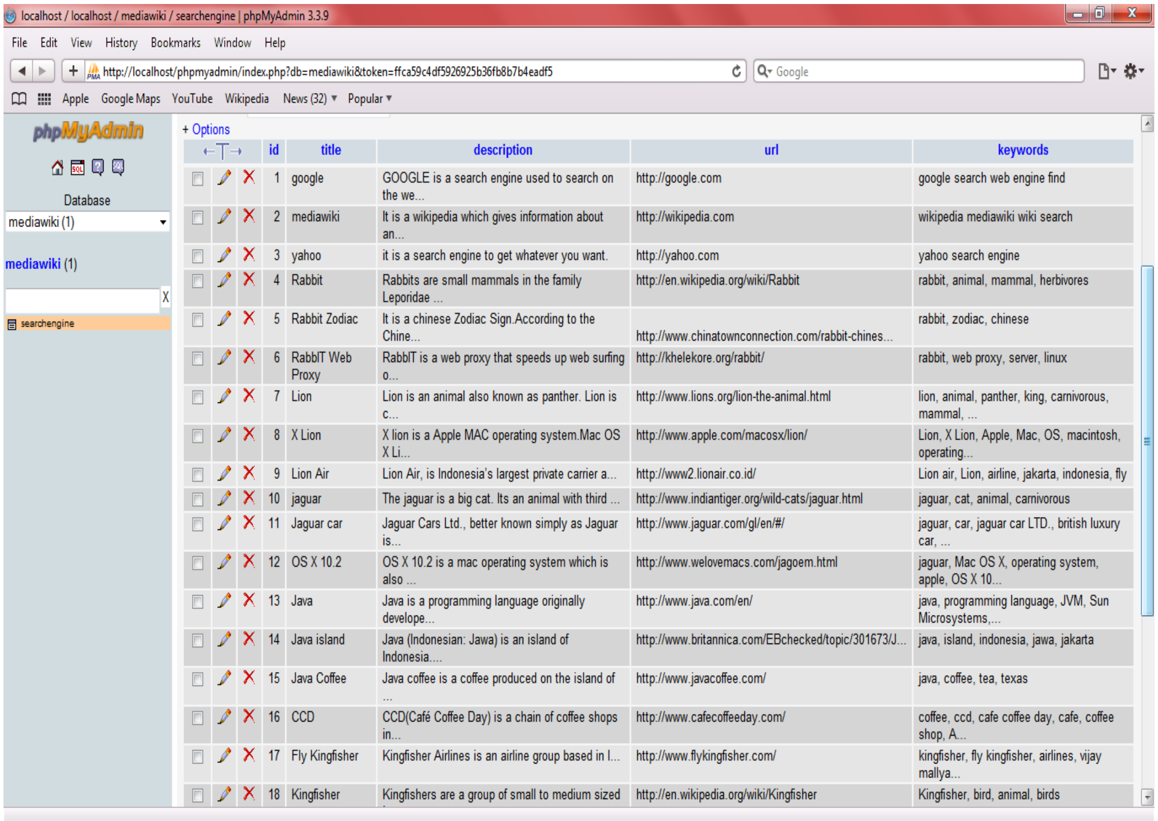

Figure 4.4: Snapshot of MySQL database

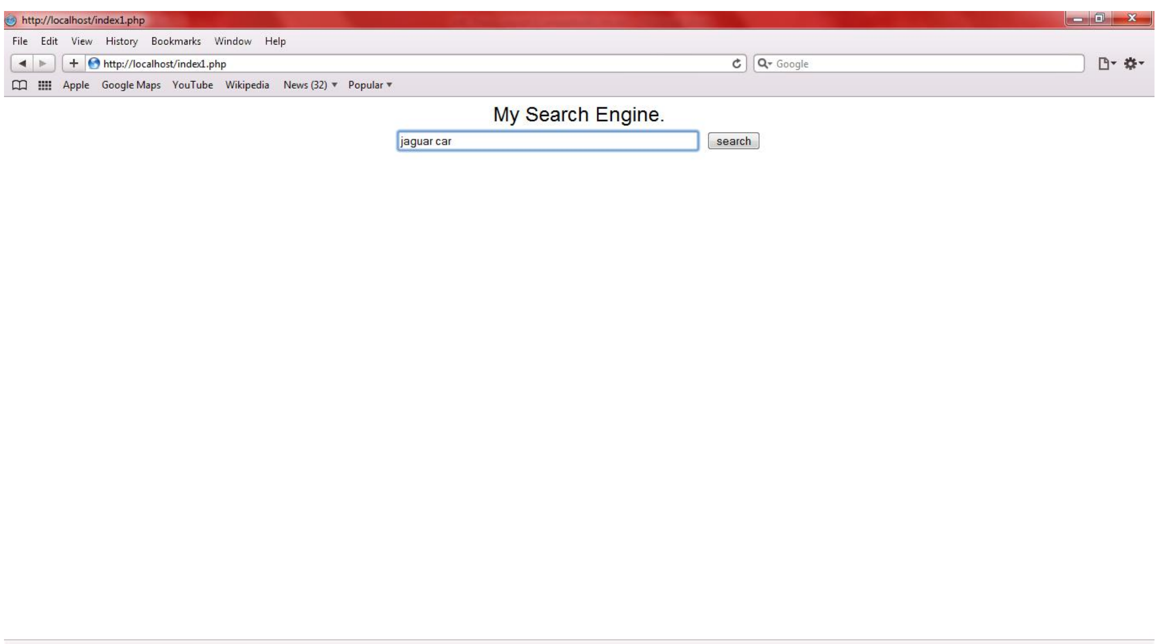

Figure 4.5: Snapshot of the Search Engine page where the query has to be entered.

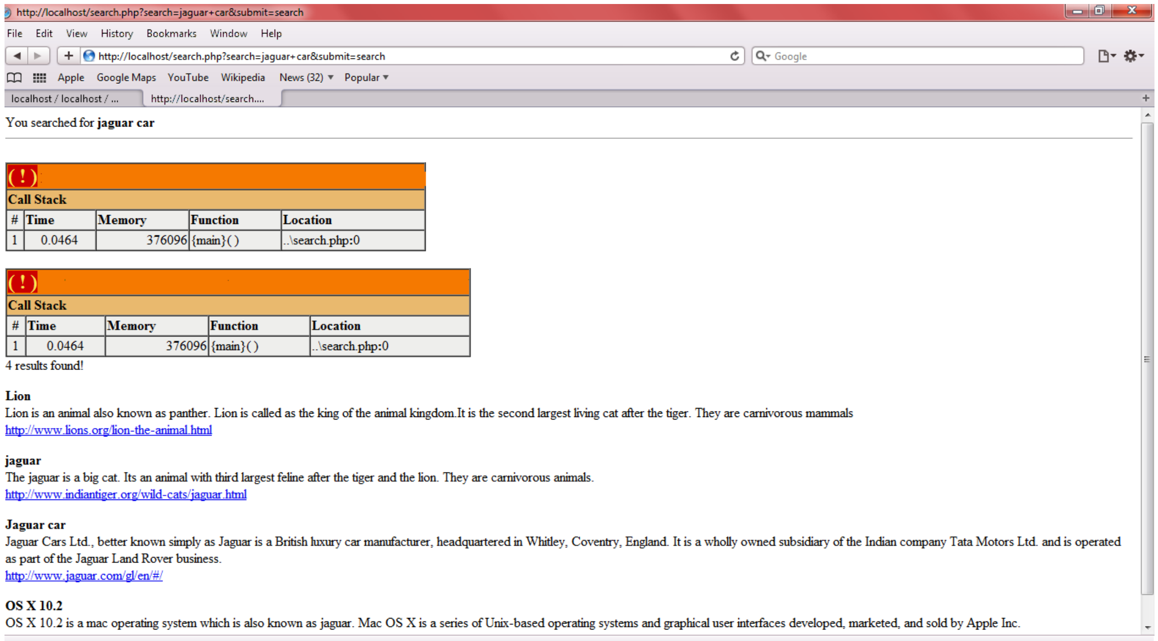

Figure 4.6: Snapshot of search retrieval page when searched for “jaguar car”

4.3 Comparison

After the semantic and syntactic information retrieval system were created, a comparison is made on the basis of precision, recall and various fundamental search facilities like symbols, keywords, wild cards, case sensitiveness etc. To measure the precision and recall 10 search queries are chosen from the database and same queries are made to run on both the environments, to get the result. The final analysis is depicted via comparison table and line graph.

4.4 Ontology based Information retrieval for learning styles of autistic people

After that based on the results of the above comparison, an “Ontology based Information retrieval for learning styles of autistic people” is created. In this proposal an ontology based prototype system for information retrieval on the Internet is described. User is interested in the focused results regarding a product with some specific characteristics. A product may have different characteristics like price, weight, size, length, color, functionality based parameters etc. It is, however, too difficult for autistic people to identify appropriate keywords due to their lack of ability to process and retain the information. Therefore, a large amount of unwanted and irrelevant data is included in the outcome. The proposal handles the search queries related to the autistic people based on their learning styles. User may type the search queries using some words or a paragraph. The objective of the work will be to find the right set of keywords from the search paragraph and retrieval of correct keywords and then finding the correct patterns or products from the web. This is based on memories of such people and their learning styles that help them find the desired result.

4.4.1 Information system for Autistic People

From the difficulties faced by the autistic people, it can be inferred that, for autistic people to remember something, they must first recall when and where the event happened, and then identify it. When they are searching for something with Web information retrieval systems, they need to recall memories of the events already occurred and use them for their search. In particular, the most important parts of their memories are memorized in images in the form of a video or a short story. Therefore, episodic memories and sequence of pictures, diagrams, movies etc. can be used to support them in their Web retrieval tasks.

It is also realized that autistic individuals are more likely to rely on only one style of learning. By observing the person, one may be able to determine and adopt his style of learning. For example, if an autistic child enjoys looking at books (e.g., picture books), watching television (with or without sound), and tends to observe people and objects carefully, then he/she may be a visual learner. Many researchers stress on the point that visualizing verbal expressions improves difficulties with communication and helps in social interaction and imagination. Verbal instructions alone are usually not a sufficient form of communication for these children. It is easier for a child with Autism to understand the teacher’s intended communications by visual facial cues, hand gestures and body language.

4.4.2 Autistic Learning style Ontology

Autistic children are motivated by imagination and consistent growth that is why computers can be thought of as the perfect assistive technology for learning. They make the child learn to function independently by putting the child in the driver’s seat. It has been proved in a research that autistic children who use computers have increased attention spans, can stay longer in their seats, have developed fine motor skills and can repeat a behaviour that was learned at school that depicts a greater ability to generalize skills across environments. From the above it is inferred that computers has provided greater benefits, so they should be an integral part of a special education student’s daily curriculum.

Moreover computer games can help autistic children in various educational areas such as learning new vocabulary, improving math skills and teaching eye-hand coordination. Educational computer games that are developed for autistic and non-autistic children are appropriate. These games provide many educational benefits for autistic children. There are a variety of single-player and multiple-player games that can improve important developmental and behavioural skills.

The "Autistic Learning style Ontology" consists of various events in their lives in the form of entities and information regarding those events as annotations. The retrieved events that occur in autistic people’s daily lives are stored as subclass information of the event class. Therefore, by asking the child when and where the event happened, and by showing images related to the events the system can infer which kind of event happened to the child. Such questions asked of the child can help them to discover the queries and get the desired results [16]. Below is a section of the “Autistic learning style Ontology” which presents various games for autistic people. Figure 4.7 gives a class view annotation of Autistic Learning style Ontology and Figure 4.8 gives the graphical representation of the ontology [18].

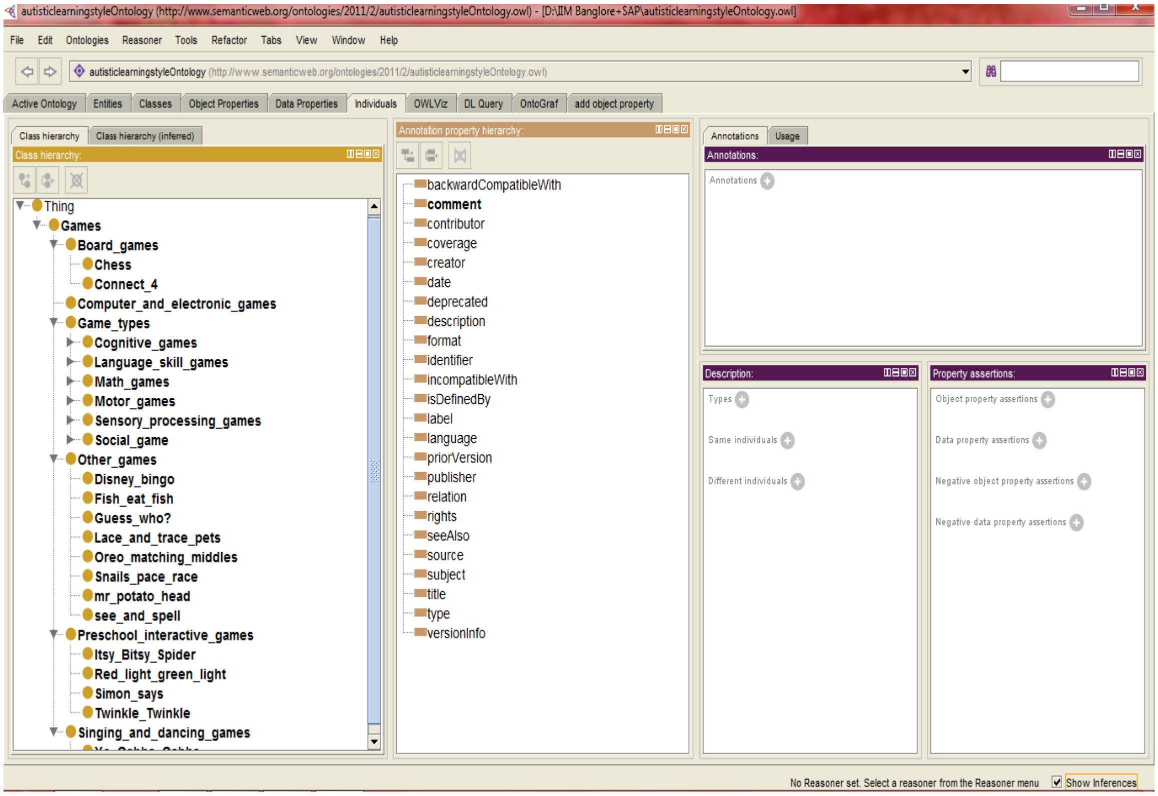

Figure 4.7: Class view Annotation for Autistic learning Style Ontology.

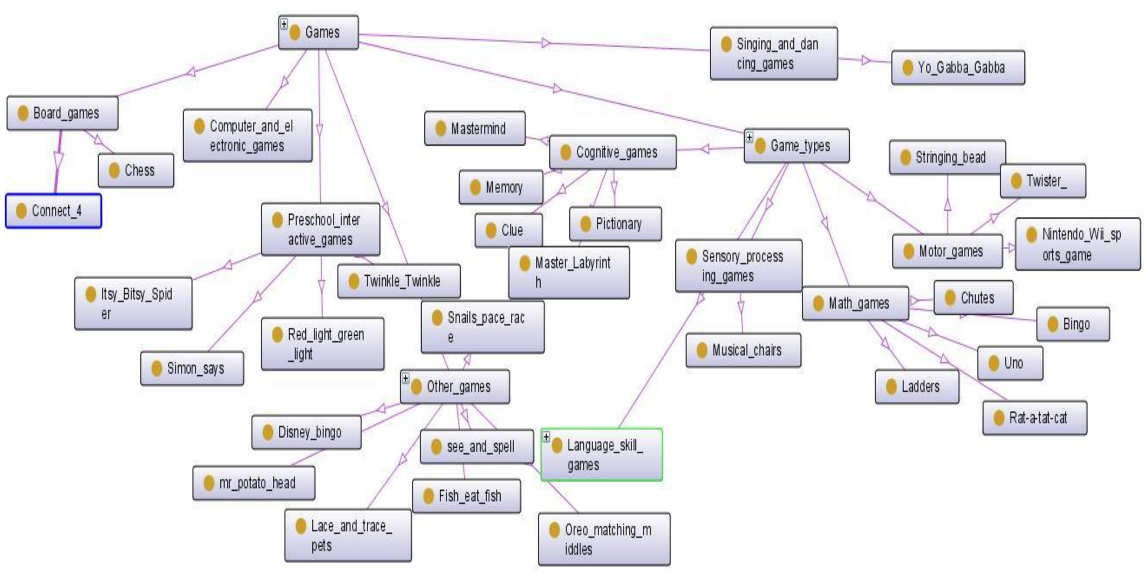

Figure 4.8: Autistic learning style Ontology graph

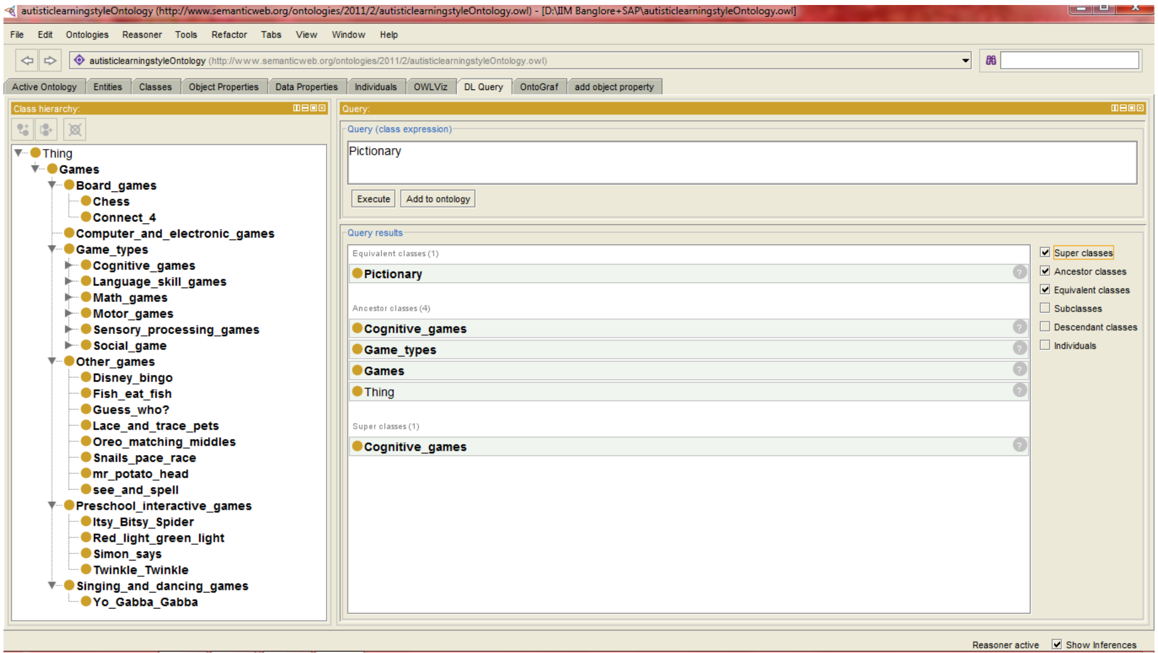

Figure 4.9: Snapshot of DL query showing the Super class, Ancestor class and Equivalent class of class expression “Pictionary”.

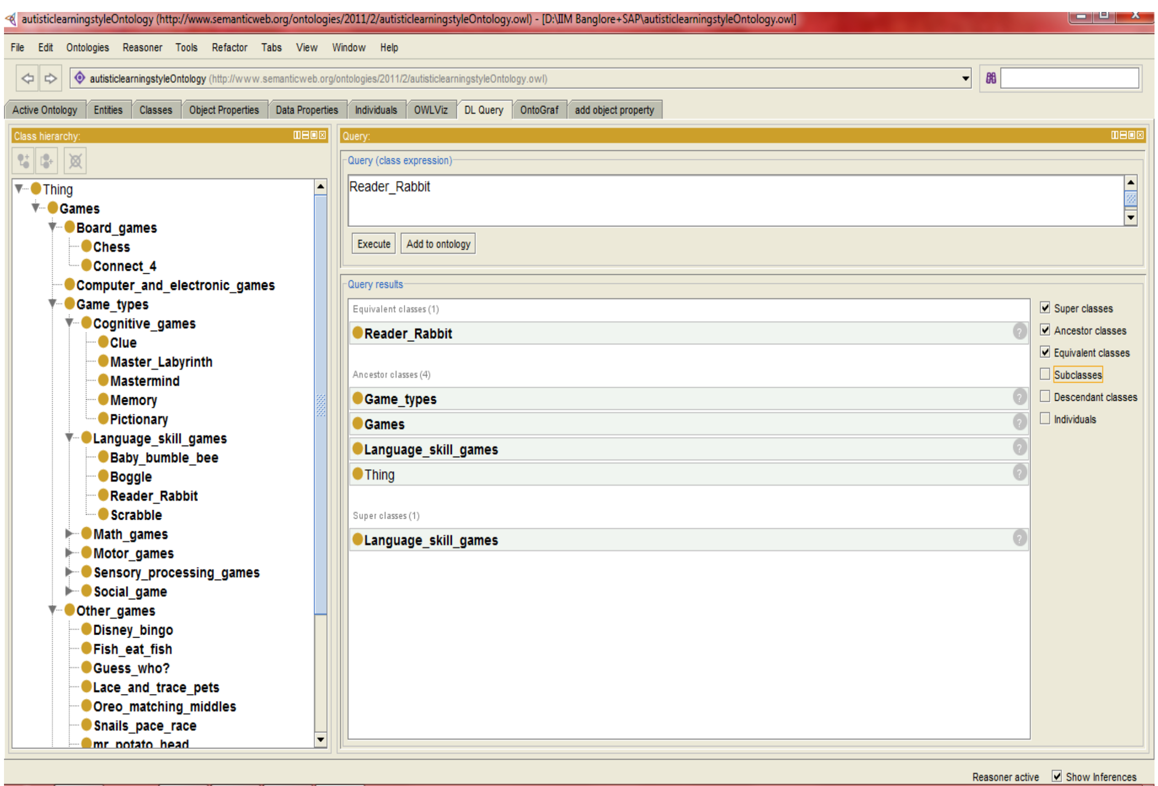

Figure 4.10: Snapshot of DL query showing the Super class, Ancestor class and Equivalent class of class expression “Reader_Rabbit”

Figure 4.9 and Figure 4.10 gives a snapshot of DL query showing the Super class, Ancestor class and Equivalent class of class expressions “Pictionary” and “Reader_Rabbit” respectively.

4.4.3 Web Retrieval process

A web retrieval process is discussed in this section. Imagine this scenario. An autistic child Bunny aged 9 wants to play a game “reader rabbit” which he saw his friend playing on his computer a day before. He doesn’t know the name of the game but know the name of the character “rabbit”. If he search that term in any search engine, the results presented are hardly helpful. There are listings for animals, Chinese zodiac sign rabbit, web proxy mixed into the results. Only after sifting through multiple listings and reading through the linked pages will he be able to find the exact thing he wants. But being an autistic child he is not able to do this.

However, in a Semantic Web-enabled environment, you could use a Semantic Web agent to search the Web for “reader rabbit”. The following are interactions between Bunny and the web Agent:

1. Bunny inputs "rabbit" as a search keyword. The resulting data include many irrelevant pages such as animal, zodiac etc.

2. Agent: "Where did you see rabbit?"

3. Bunny inputs "game".

4. The system collects a set of character names in popular games on the Web.

5. Agent: "You are searching for Reader Rabbit in which a rabbit with blue and red striped sweater plays?

6. Bunny: inputs "Yes."

7. Agent: The system dynamically finds the keyword "Reader rabbit" in the “Autistic learning style Ontology." The subclass hierarchy consists of "Games," "Game types," "Language skill games" and "Reader Rabbit".

Agent: "Would you like to play the game Reader Rabbit?"

8. Bunny: inputs "Yes"

9. The agent starts the game on the computer and tells him about the rules of the game.

Chapter- 5
EXPERIMENTAL RESULTS

5.1 Comparison Based On Fundamental Search Facilities

Table 5.1 gives the comparison of semantic and syntactic information retrieval system on the basis of various fundamental search facilities like symbol used, keywords used in the search queries, phrases, wildcards, prefixes etc.

Table 5.1: Comparison of semantic and syntactic Information Retrieval system.

Information retrieval system/ Properties	**Semantic Information Retrieval system**	**Syntactic Information Retrieval system**
Symbol	∃, ∋, ≥, =, ≤, ∀	+, -, ( )
Keywords	some, value, min, exactly, max, only	AND, OR, ANDNOT
Phrase	“ ”, [ ]	“ ”
Wildcards	*, ?, $	(*) whole word wildcard
Case sensitive	YES	NO
Prefixes	length, maxLength, minLength, totalDigits, fractionDigits	filetype, inurl

.

5.2 Estimation of Precision and Recall

To measure information retrieval effectiveness in the standard way, we need a test collection consisting of three things [1]:

1. A document collection i.e. a database from which the search is to be performed.
2. Information needs, expressible as queries.
3. A binary assessment of either relevant or non-relevant for each query-document pair.

To measure the effectiveness two parameters are defined: Precision and recall.

Precision (P) is the fraction of retrieved documents that are relevant

Precision = #(relevant items retrieved) / #(retrieved items)

= P(relevant|retrieved)

= P(sum/#)

Recall (R) is the fraction of relevant documents that are retrieved

Recall = #(relevant items retrieved) / #(relevant items)

= P(retrieved|relevant)

=P(num/#)

To measure the precision and recall [19], [20], both the semantic and syntactic information retrieval systems are tested for 10 queries and based on the results which are retrieved, the estimation is made. The search- items (queries) on which estimation are done:

#1: Operating system

#2: Jaguar car

#3: Web Proxy

#4: Fly Kingfisher

#5: Rabbit Zodiac

#6: Coffee

#7: X lion

#8: OS X 10.2

#9: Prolog language

#10: Carnivorous animal

Table 5.2: Estimation of Precision

Search Item	**Syntactic Information retrieval system**		**Semantic Information retrieval system**	
	P(sum/#)	**P**	**P(sum/#)**	**P**
#1	2.0/3.0	0.67	2.0/3.0	0.67
#2	1.0/4.0	0.25	1.0/1.0	1
#3	1.0/2.0	0.5	1.5/2.0	0.75
#4	1.0/3.0	0.34	1.0/5.0	0.2
#5	1.0/3.0	0.34	1.0/1.0	1
#6	1.0/2.0	0.5	4.0/5.0	0.8
#7	3.0/8.0	0.375	1.0/1.0	1
#8	1.0/3.0	0.34	1.0/2.0	0.5
#9	1.0/3.0	0.34	1.0/1.0	1
#10	2.0/4.0	0.5	5.0/8.0	0.625

Mean P	N/A	0.416	N/A	0.734

Table 5.2 gives the estimation of precision for both semantic and syntactic search engine based on various search items. These search items are run on both the environments to get the results. From the Mean P, it can be seen that mean precision for semantic retrieval system is much higher as compared to that of syntactic search system.

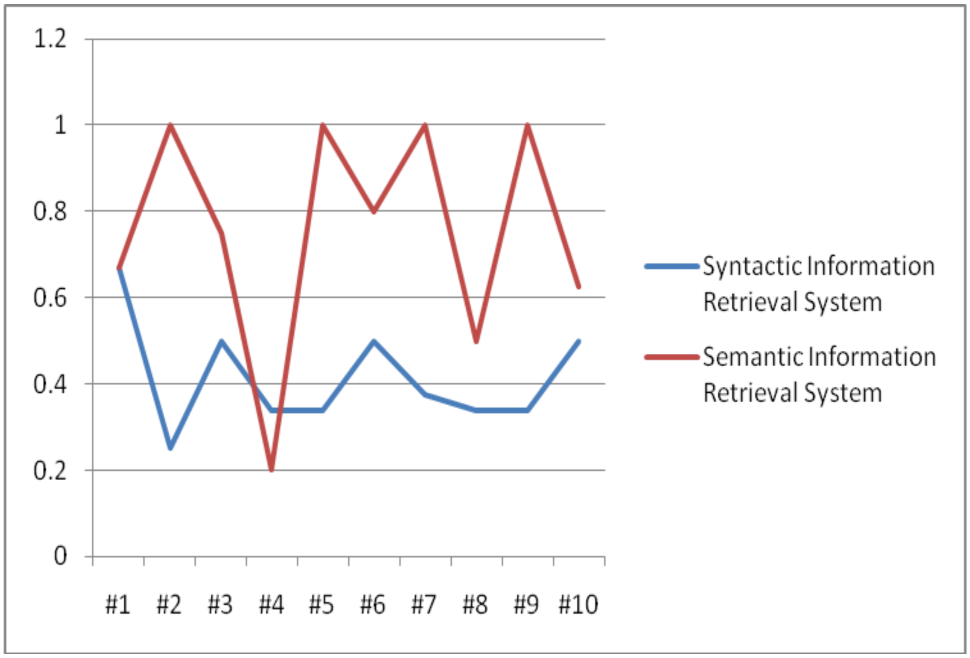

Figure 5.1: Comparison on the basis of precision

Mean precision

- Syntactic Information Retrieval system= 0.416
- Semantic Information Retrieval system= 0.734

Figure 5.1 gives the graphical representation of the above table. From the above table a graph is plotted which gives the comparison of two environments on the basis of precision. From the graph it can be inferred that the semantic information retrieval system has a higher precision for the same search items as compared to the syntactic information retrieval system.

Table 5.3: Estimation of Recall

Search Item	**Syntactic Information retrieval system**		**Semantic Information retrieval system**	
	P(num/#)	**R**	**P(num/#)**	**R**
#1	2.0/3.0	0.67	3.0/3.0	1
#2	1.0/2.0	0.5	1.0/1.0	1
#3	1.0/2.0	0.5	0/2.0	0
#4	1.0/2.0	0.5	1.0/2.0	0.2
#5	1.0/1.0	1	1.0/1.0	1
#6	2.0/5.0	0.4	4.0/4.0	1
#7	3.0/3.0	1	1.0/3.0	0.34
#8	1.0/1.0	1	0.5/1.0	0.5
#9	2.0/3.0	0.67	1.0/3.0	0.34
#10	2.0/5.0	0.4	5.0/5.0	1
Mean R	N/A	0.664	N/A	0.638

Table 5.3 gives the estimation of recall for both semantic and syntactic search environments when run on same search items. As can be seen from the above table, for

some of the search items, semantic retrieval system performs better and for other semantic retrieval system performs better. So a clear distinction is not made between the two. The mean recall for the search environments is approximately the same.

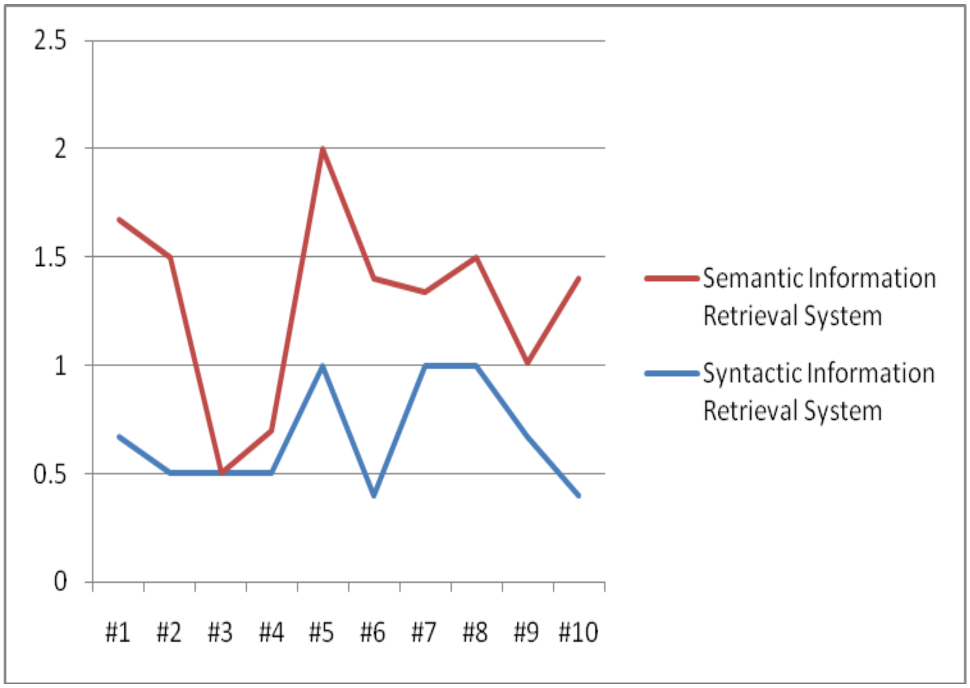

Figure 5.2: Comparison on the basis of recall

Mean recall

- Syntactic Information Retrieval system= 0.664
- Semantic Information Retrieval system= 0.638

Figure 5.2 gives the graphical representation of the above table. A graph is plotted between recall and the search items to give a comparison of the two search environments. As can be seen from the graph, semantic information retrieval system, shows a large diversity in the recall ratio for different search items i.e. for some search items the recall rate is very high and for others, it is nearly zero. Whereas for syntactic search retrieval system a constant recall rate can be seen. The graph shows a consistent rate and is not fluctuating.

Though the mean recall rate is approximately the same for both the semantic and syntactic retrieval systems, it can be inferred that syntactic retrieval system shows a more consistent recall rate as compared to semantic retrieval system, which has a fluctuating recall rate.

Chapter 6

CONCLUSION & FUTURE SCOPE

6.1 Conclusion

A competent IR system must include the fundamental search facilities that are familiar to the user, which can include boolean logic, phrase searching, wild cards and use of prefixes [21], [22]. Because the searching capabilities of IR system ultimately determine its performance, absence of such basic functions can severely affect the working of the search tool [25]. As we can see in Table 5.1, a comparison is done based on these search facilities. Both semantic and syntactic information retrieval system uses various search facilities but popularity of syntactic web is more as compared to semantic web as the former is widely used and accepted whereas the semantic web is new.

Moreover retrieval performance is traditionally evaluated on two parameters: precision and recall. While these two variables can be quantitatively measured, extra care should be exercised when judging the relevance of retrieved items and estimating the total number of documents that are relevant to a specific topic in the retrieval system [25]. Figure 5.1 gives the comparison of precision for both semantic and syntactic information retrieval system. Clearly it can be seen from Table 5.2 that semantic information retrieval system have mean precision of 0.734 which is much higher than that of syntactic information retrieval system which have a mean precision of 0.416 only. So it can be said that ratio of relevant items retrieved to total items retrieved for a search query is better in case of semantic information retrieval system.

Similarly Figure 5.2 gives the comparison on the basis of recall. As can be seen from Table 5.3 the mean recall for both semantic and syntactic information retrieval system is almost the same. So the ratio of number of relevant items retrieved to the total number of relevant items present is almost same for both the retrieval system. Moreover from Figure

5.2, it can be inferred that the syntactic retrieval system has a more consistent recall rate as compared to syntactic search retrieval system which has a fluctuating recall rate.

Then based on the above results and the performance of both semantic and syntactic Information Retrieval System, it is inferred that the Semantic Web provides a futuristic approach to the web. In this information is given explicit meaning, making it easier for machines to automatically process and integrate information available on the Web.

Using this idea an Ontology for autistic people is designed which has an advantage over the existing ontology. The main problem with the existing ontology is that it does not provide the latest information because it is static and fixed, no new data can be added to it. Whereas this “autistic learning style Ontology”, can be modified and enhanced taking the latest information from various sources related to the autistic people. E.g. Blog’s written by mothers, teachers of autistic children and from the web. This ontology is to help autistic people retrieve information about the various events happening in their lives and their learning styles.

When children want to search for something, they use memories and sequence of images in the form of a video or a short story to express what they want. To identify the right query, their unique and special ways to express things needs to be understood. By using this “autistic learning style ontology, it is expected to better understand their needs and learning styles.

6.2 Future Scope

- This autistic learning style ontology can be improved by adding more information and can be used to help the autistic children on the web. So that it is easy for the autistic people to retrieve information and use them in their daily lifestyle.
- This ontology can be generalized to cover other disabilities i.e. physically disabled, mentally challenged etc.

- Moreover a clearer picture of the recall ratio can be made out by implementing the retrieval systems on a larger warehouse as it can be seen in this case that recall ratio is approximately the same for both semantic and syntactic information retrieval systems.

- A search engine can also be implemented by combining the properties of various information retrieval systems so that better results are achieved.

REFERENCES

1. C.D. Manning, P. Raghavan, H. Schütze. "An Introduction to information retrieval", Cambridge University Press Cambridge, England, pp. 26- 569, Apr 1, 2009.

2. Guo-Qiang Zhang, Adam D. Troy, and Keith Bourgoin. "Bootstrapping Ontology Learning for Information Retrieval Using Formal Concept Analysis and Information Anchors", Department of Electrical Engineering and Computer Science Case Western Reserve University Cleveland, Ohio 44106, U.S.A, pp.1-14, 2008.

3. Williams D. "Autism and Sensing: The Unlost Instinct." Jessica Kingsley Publishers, 1998.

4. M. Horridge, H. Knublauch, A. Rector, R. Stevens, C. Wroe. "A Practical Guide To Building OWL Ontologies Using The Prot´eg´e-OWL Plugin and CO-ODE Tools Edition 1.0", The University Of Manchester, 2004.

5. World Wide Web Consortium. "OWL Web Ontology Language Semantics and Abstract Syntax". W3C Recommendation, 10 Feb, 2004.

6. Poole J. B., Skymcilvain E., Jackson L., and Singer Y. (Book) "Education for an Information Age: Teaching in the Computerized Classroom", 5th Edition, pp. 401, 2005.

7. Microsoft Corporation History of Microsoft's Commitment to Accessibility (On-line). Web: http://www.microsoft.com/enable/microsoft/history.aspx, 2004.

8. World Wide Web Consortium Web Content Accessibility Guidelines 2.0 (On-line). http://www.w3.org/TR/WCAG20/checklist.html, 2005.

9. A. Basu, S. Sarkar, K. Chakraborty, S. Bhattacharya, and M. Choudhary "Vernacular Education and Communication Tool for the People with Multiple Disabilities." In the Digital Library of the 2nd International Conference on Open Collaborative Design for Sustainable Innovation (On-line), Bangalore, India. Web: http://www.thinkcycle.org/tc-papers, 2002.

10. Griswold, D. E., Barnhill, G. P., Myles, B. S., Hagiwara, T. and Simpson, R. L. "Asperger syndrome and academic achievement." Focus on Autism and Other Developmental Disabilities, summer, 17, 2, pp. 94, 2002.

11. K. Yılmaz, U. Ozlem, Serdar E. Guner, Bal K. "An NLP-Based Assistive Tool For Autistic And Mentally Retarded Children: An Initial Attempt", Trakya University, Faculty of Engineering and Architecture, Department of Computer Engineering 22100 Edirne Turkey, pp.1-8.

12. Augmentative and Alternative Communication (AAC) Connecting Young Kids (YAACK). Web: http://aac.unl.edu/yaack/, 1999.

13. Williams D. "Autism: An Inside-Out Approach." Jessica Kingsley Publishers, 1996.

14. Gradin T. "Thinking In Pictures." Vintage Press, 2006.

15. C. Cesarano, A. d'Acierno, A. Picariello. "An Intelligent Search Agent System for Semantic Information Retrieval on the Internet". WIDM'03, New Orleans, Louisiana, USA. Nov 7–8, 2003.

16. H. Knublauch, M. A. Musen, A. L. Rector. Medical Informatics Group, "Editing Description Logic Ontologies with the Protege OWL Plugin", Stanford University and University of Manchester, pp. 1- 9.

17. World Wide Web consortium Internet: http://www.w3.org/2001, 2001.

18. V. David, F. Miriam, C. Pablo. “An Ontology Based Information Retrieval Model” Universidad Autonoma de Madrid, pp. 1-15.

19. J. Bar-Ilan. “On the overlap, the precision and estimated recall of search engines: A case study of the query "Erdos"”. Scientometrics, 42 (2), 207-208, 1998.

20. S. Clarke, P. Willett. “Estimating the recall performance of search engines”. ASLIB Proceedings, 49 (7), pp. 184-189, 1997.

21. W. Ding, G. Marchionini. “A comparative study of the Web search service performance”. In: Proceedings of the ASIS 1996 Annual Conference, pp.136-142, Oct 1996.

22. C. Oppenheiem, A. Moris, C. Mcknight, S. Lowley. “The evaluation of WWW search engines”. Journal of documentation, 56 (2), pp.190-211, 2000.

23. H. Sumiyoshi, I. Yamada, Y. Murasaki, Y.B. Kim, N. Yagi and M. Shibata, "Agent Search System for A New Interactive Education Broadcast Service", NHK STRL R&D No.84, Mar, 2004.

24. Eero HyvÄonen, Avril Styrman, and Samppa Saarela. “Ontology-Based Image Retrieval”, University of Helsinki, Department of Computer Science, pp.1-13.

25. H. Chu, M. Rosenthal. “Search engines for the World Wide Web: a comparative study and evaluation methodology”, Proceedings of the ASIS 1996 Annual Conference. [online] Available: http://www.asis.org/annual96/ElectronicProceedings/chu.html. October, 33. 127-35. Retrieved August 19, 2003.

26. Alan R. Aronson, Thomas C. Rindflesch, Allen C. Browne. “Exploiting a Large Thesaurus for Information Retrieval” National Library of Medicine 8600 Rockville Pike Bethesda, pp.1-19, 2 Mar, 2010.

LIST OF PUBLICATIONS

1. Sanchika Gupta, Dr. Deepak Garg "Ontology based Information retrieval for learning styles of autistic people" at ICHPAGC 2011, Springer Press Germany Communications in Computer and Information Science series- 169. **[Published]**.

2. Sanchika Gupta, Dr. Deepak Garg "Comparison of Semantic and Syntactic Information Retrieval System on the basis of Precision and Recall" at International Journal of Data Engineering (IJDE), Volume (2) : Issue (3), 2011 [**Communicated**].