Comparison of Semantic and Syntactic Information Retrieval System on the Basis of Precision and Recall

Sanchika Gupta sanchigr8@gmail.com
Student/Computer Sc. & Engg.
Thapar University
Patiala, 147004, India

Dr. Deepak Garg dgarg@thapar.edu
Faculty/Computer Sc. & Engg./Asstt. professor
Thapar University
Patiala, 147004, India

Abstract

In this paper information retrieval system for local databases are discussed. 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. 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.

Keywords: Information Retrieval, Precision, Recall, Semantic, Syntactic.

1. INTRODUCTION

Information Retrieval (IR) is the study of systems for searching, retrieving, clustering and classifying the data, particularly text or other unstructured forms. IR is finding material of an unstructured nature that satisfies an information need from within large storage usually from the computers [1]. IR is also used to facilitate semi structured search, clustering of documents based on their contents and classification of data. 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. After the database is created a query is entered in the search space. A query is request for information from a database. These are formal statements for satisfying information needs. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy and accuracy. In general, a query is a form of questioning, in a line of inquiry. The style and format of querying might be different for both syntactic and semantic search engine.

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”.
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 World Wide Web 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.

In this paper information retrieval system for local databases are discussed. 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. 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.

2. SEMANTIC INFORMATION RETRIEVAL SYSTEM

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 [13]. The GUI serves as an interface between the user and the system. OWL (ontology web language) is the language used for developing ontologies. OWL Properties represent relationships. There are three types of properties-

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

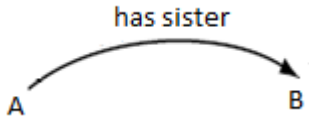

FIGURE 1: An object property linking the individual A to individual B.

- Datatype properties

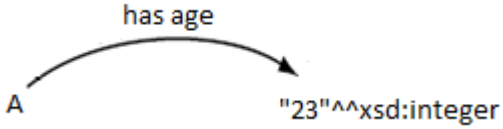

FIGURE 2: A datatype property linking the individual A to data literal '23', which is a type of integer.

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

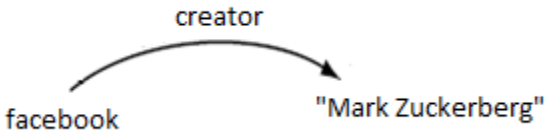

FIGURE 3: An annotation property, linking the class 'facebook' to the data literal (string) "Mark Zuckerberg".

OWL	DL Symbol	Manchester OWL Syntax Keyword	Example
someValuesFrom	∃	some	hasChild some Man
allValuesFrom	∀	only	hasSibling only man
hasValue	∋	value	hasCountryOfOrigin value England
minCardinality	≥	min	hasChild min 3
cardinality	=	exactly	hasChild exactly 3
maxCardinality	≤	max	hasChild max 3

TABLE 1: Description logic symbols and the corresponding English language keywords [4].

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 favourite 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 [15]. The semantic web agent helps you to find the required car and also tells you about its features, functions, price and other available options. FIGURE 4 presents the media wiki ontology graph which consists of related information.

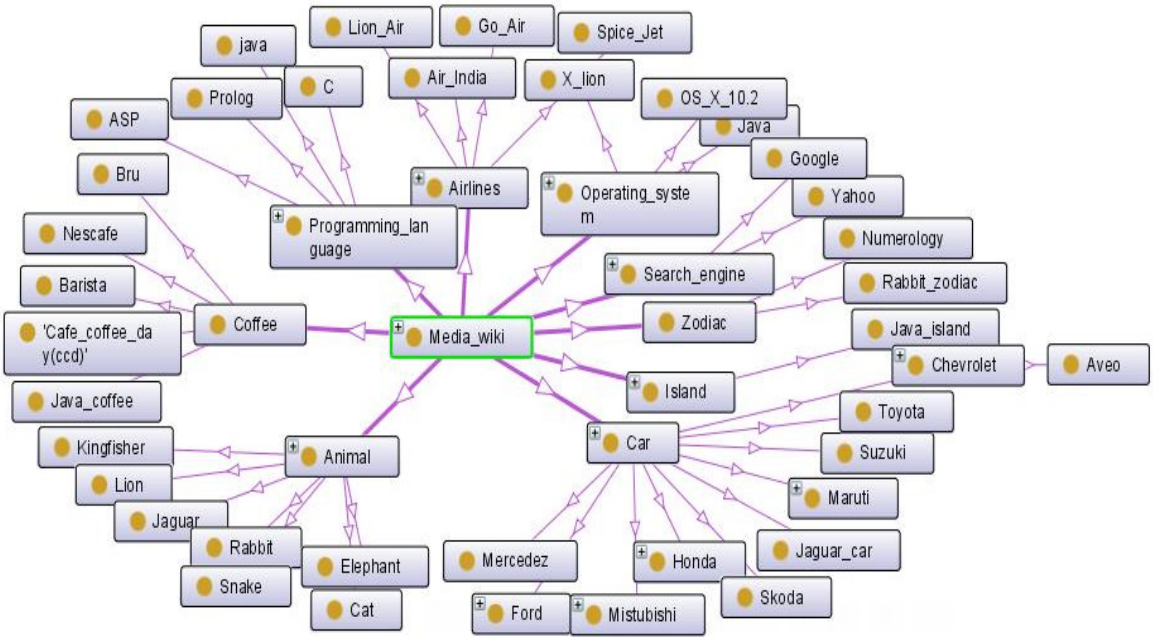

FIGURE 4: Mediawiki Ontology graph

3. SYNTACTIC INFORMATION RETRIEVAL SYSTEM

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 users 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 [12].

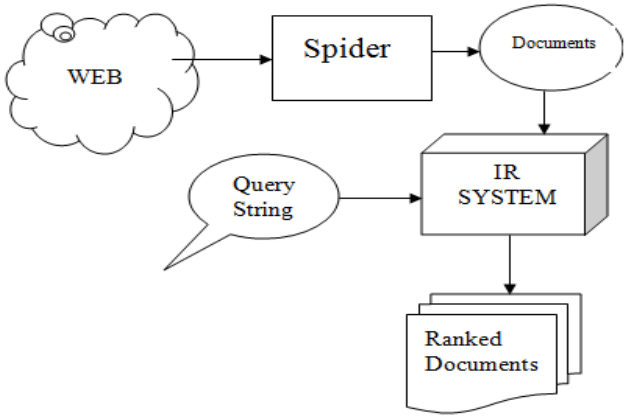

FIGURE 5: Syntactic Information Retrieval System.

FIGURE 5: depicts the syntactic information retrieval system. It consists of a spider which is a computer program that browses the web in a 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 [14]. 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.

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. For ex. 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.

4. COMPARISON BASED ON FUNDAMENTAL SEARCH FACILITIES

This table 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.

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

TABLE 2: Comparison of semantic and syntactic Information Retrieval system.

.

5. 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, both the semantic and syntactic information retrieval systems are tested for 5 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

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
Mean P	N/A	0.42	N/A	0.72

TABLE 3: Estimation of Precision

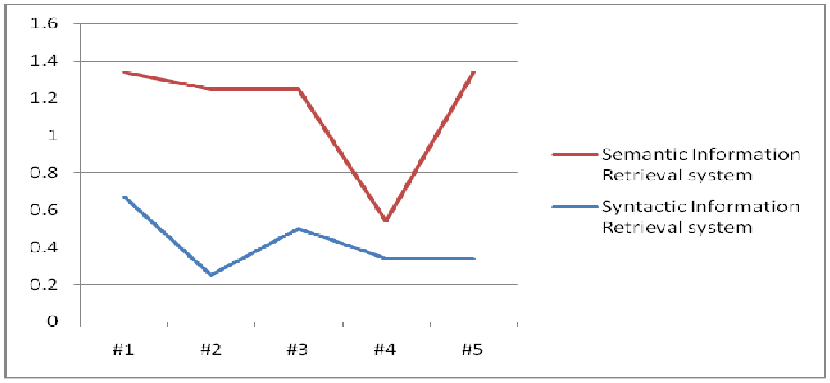

FIGURE 6: Comparison on the basis of precision

Mean precision

- Syntactic Information Retrieval system= 0.42
- Semantic Information Retrieval system= 0.72

Figure 6 gives the graphical representation of the above table. From Table 3, 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.

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
Mean R	N/A	0.634	N/A	0.64

Table 4: Estimation of Recall

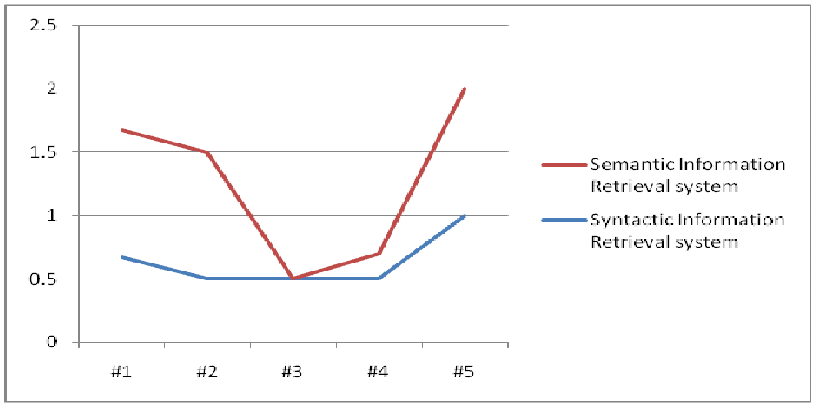

FIGURE 7: Comparison on the basis of recall

Mean recall

- Syntactic Information Retrieval system= 0.634
- Semantic Information Retrieval system= 0.64

Figure 7 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.

6. CONCLUSION

A competent Information Retrieval system must include the fundamental search facilities that users are familiar with, which include Boolean logic symbols, phrase searching, wild cards and use of prefixes. Because the searching capabilities of Information Retrieval system ultimately determine its performance, absence of these basic functions will severely handicap the search tool [8]. As we can see in Table 2, 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.

Retrieval performance is traditionally evaluated on two parameters: precision and recall. While the two variables can all be quantitatively measured, extra caution should be exercised when one judges the relevance of retrieved items and estimates the total number of documents relevant to a specific topic in the retrieval system [8]. FIGURE 6 gives the comparison of precision for both semantic and syntactic information retrieval system. Clearly it can be seen that semantic information retrieval system have mean precision of 0.72 which is much higher than that of syntactic information retrieval system which have a mean precision of 0.42 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 7 gives the comparison on the basis of recall. As can be seen from the table 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 7, 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.

7. REFERENCES

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

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

[3] 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.

[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 Internet: http://www.w3.org/2001, 2001.

[6] V. David, F. Miriam, C. Pablo. "An Ontology Based Information Retrieval Model" Universidad Autonoma de Madrid.

[7] 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.

[8] H. Chu, M. Rosenthal. (1996). "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.

[9] S. Clarke, P. Willett. "Estimating the recall performance of search engines". ASLIB Proceedings, 49 (7), pp. 184-189, 1997.

[10] W. Ding, G. Marchionini. "A comparative study of the Web search service performance". In: Proceedings of the ASIS 1996 Annual Conference, Oct 1996, pp.136-142.

[11] C. Oppenheiem, A. Moris, C. Mcknight, S. Lowley. "The evaluation of WWW search engines". Journal of documentation, 56 (2), pp.190-211, 2000.

[12] 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.

[13] E. HyvÄonen, A. Styrman, S. Saarela. "Ontology-Based Image Retrieval", University of Helsinki, Department of Computer Science, pp.1-13.

[14] 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.

[15] 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.