Semantic News Finder : A Semantic Retrieval from News Items

M.Thangaraj
Associate Professor
Dept. of Computer Science
Madurai Kamaraj University
Maduraii, TN, India.

G.Sujatha
Associate Professor
Dept. of Computer Science
Sri Meenakshi Govt. Arts
College for Women, Madurai.

ABSTRACT

In this paper an Ontology based semantic search approach is presented for News domain. The proposed Semantic News Finder retrieves information from News websites using the predefined News ontology. Since there is a big variety of news sources, it would be useful to provide an efficient method to organize and retrieve the relevant results. This work includes a new method for semantic query construction, creation of semantic news knowledge base, computation of similarity between expanded query and the information content. The system is implemented by using the technologies in semantic web and the performance is evaluated with traditional systems. The proposed architecture incorporates semantic annotation technology, ontology evolution, term extraction and indexing resources. The evolution of the system has produced very promising results.

Keywords

Semantic Search, News Ontology, Information Retrieval, Semantic Web, Semantic Query.

1. INTRODUCTION

1.1 Motivation

Today's web is a major source of information, and the richness of it is largely underexploited. Indeed, if its gigantic unanimous, it is difficult from its ability to meet our information needs. The question today for the search engines is not how many pages are retrieved? But, how many relevant web pages will be retrieved? A lot of time is lost in looking for user needs in the pages retrieved by the search engines, and often the user are forced to change their search queries. These systems use a centralized database for indexing information. They are based on queries from simple keywords. The recall rate is high, but the accuracy is low. This is due to the disambiguation, wrong context, the use of different words, more specific words, or more general. These systems rarely take into consideration the semantic content of the document to the index. The approach allows taking into consideration the semantics of the document focuses on techniques of information retrieval based on ontologies. For these type of systems, documents are indexed according to the ontology concepts.

The huge increase in the amount and complexity of reachable information in the WWW caused an excessive demand for tools and techniques that can handle data semantically. The current practice in IR mostly relies on keyword based search over full-text data, which is modelled as a bag-of-words. The fast spread of the internet facilitated new exchange and resulted in a dramatic increase in the number of available news sources. This in turn led to an increased volume of news items that an average recipient received every day, resulting in the abundance of information available for the consumers. However such a model misses the actual semantic information of the text.

In order to deal with this issue, ontologies are proposed for known representation which are nowadays the backbone of semantic web applications. Both the IE and IR processes can bene_t from such meta data, which gives semantics to plain text. The idea of semantic web is not to make sure that computers can understand human language or operating in natural language, it is not artificial intelligence allowing the web to think, but simply to group the information in a useful way, as a huge database, where everything is written in a structured manner.

This motivates the development of Semantic News Finder (SNF). This proposed framework handles the user query and News articles based on News ontology with an improved algorithm for indexing and ranking.

1.2 Structure of the Paper

The rest of the paper is organized as follows. The Related work is presented in section 2.The working mechanism of the proposed architecture is explained in section 3. In section 4 Performance evaluation is given and the Conclusion is presented in section 5.

2. RELATED WORK

The information society technologies also shows great interest in semantics in the News industry by funding several projects like the NEWS project [15] , which contributed by providing tools for Semantic - Based analysis, personalization and delivery of knowledge from news-wires. Another ongoing research project in the same area is SYNC# [12] which provides a method to structure news items based on a combination of linguistic and statical processing. A database query is somewhat different from semantic query and therefore its processing[16][17].

Text classification is another research area that News industry has great interest in. The classification of News articles under a predefined set of codes can be a very difficult task to perform manually. To address this issue a method is proposed in [10] using machine learning techniques.

Focused crawling approaches [[2] [3] [8]] are an alternative which aims at increasing precision. They provide search engines dedicated to a specific information domain. However, such systems usually rely on learning techniques that are difficult to apply and involve a cumbersome training phase.

Semantic search [[4]-[6],[5]] is a current trend to cope with the precision of the results. The various semantic search approaches are analysed in [13]. Based on the W3C standards and recommendations, search systems have been developed for closed worlds of knowledge for a specific domain [11].An improved architecture for semantic information retrieval [14] is presented that provides a method for semantic querying and retrieval by an improved ranking algorithm.

Ontology based meta data extraction is the latest approach. The World News Finder [9] presents a semantic search of news articles using World News Ontology. The ontology is used to extract meta data from news articles and perform semantic search on them. This system is in lack of self-adopting learning system to provide a way to automatically produce the set of values that should be assigned to the parameters in order to maximize its efficiency with a less time-consuming effort.

Ontology-based Personalised Context-aware Recommendations of News items [7] is designed with news contents and user preferences in terms of concepts appearing in a set of domain ontologies. The re-ranking approach of retrieved web documents based on the relevancy measure is presented in [1]. The literature survey revealed that current studies on keyword-based semantic searching are not mature enough either they are not scalable to large knowledge bases or they cannot capture all the semantics in the queries. Our main contribution is to fill this gap by implementing a Semantic News Finder using the semantic indexing approach with an efficient ranking algorithm.

The study presented in this paper can be extended to other domains as well as by modifying the current ontology and the information extraction module.

3. PROPOSED ARCHITECTURE

A high level view of the overall system is given in Fig. 1. The overall scenario utilizes ontology as the basis to transform query and the semantic features in the context repository into semantic pattern so as to identify the corresponding contents by Semantic News Retriever. The result of this retriever is ranked with an improved ranking algorithm based on the semantic entities. A more detailed description of the Semantic News Finder (SNF) architecture is presented in Fig.2.

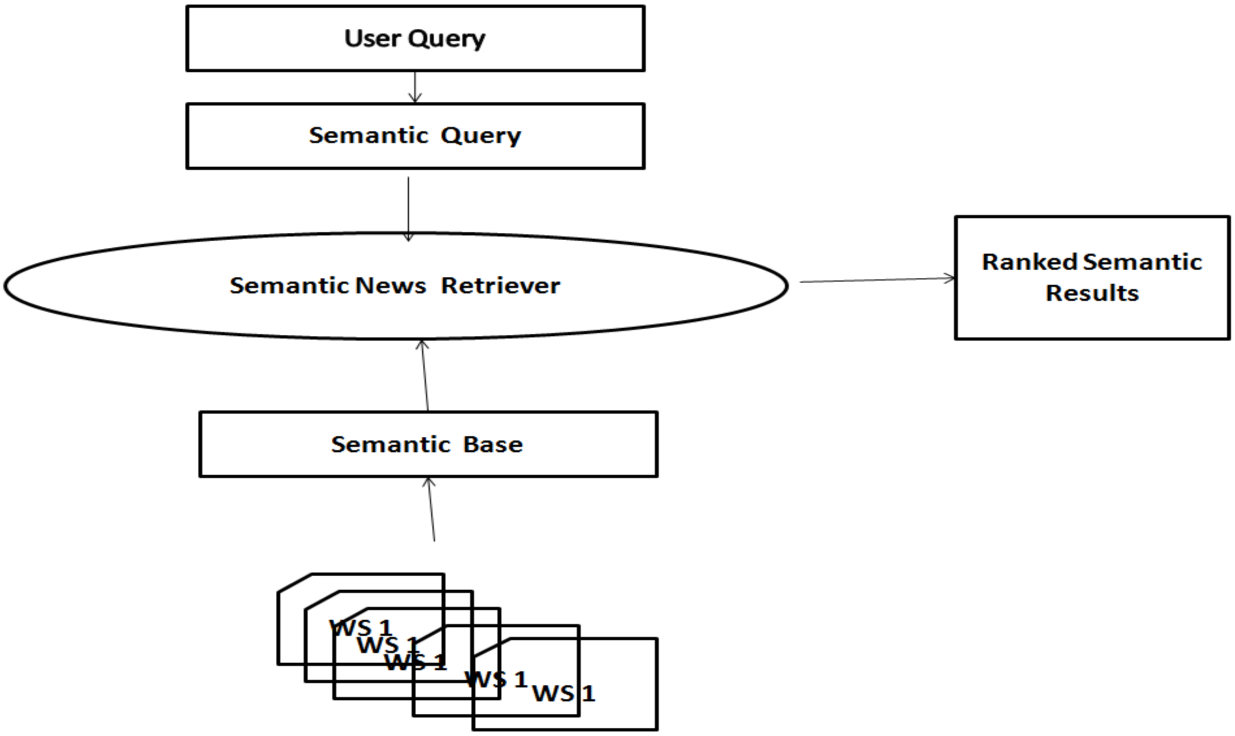


Figure 1: Overall view of the System

This system is composed of three main layers namely Data Layer, Logic Layer and Application Layer. In short the system comprises two separate phases: i)Creation of Semantic News Repository ii)Discovery of relevant results. The data layer collects News articles from various news websites. The logic layer incorporates four main components as Semantic News Annotator and Indexer, Semantic Query Converter, Semantic News Retriever and News Ranker. At the Application layer an interactive user interface is presented to query the system. Semantic search is performed based on the new search algorithm and the relevant results are presented to the end user based on the developed ranking algorithm. Next these components are described in detail.

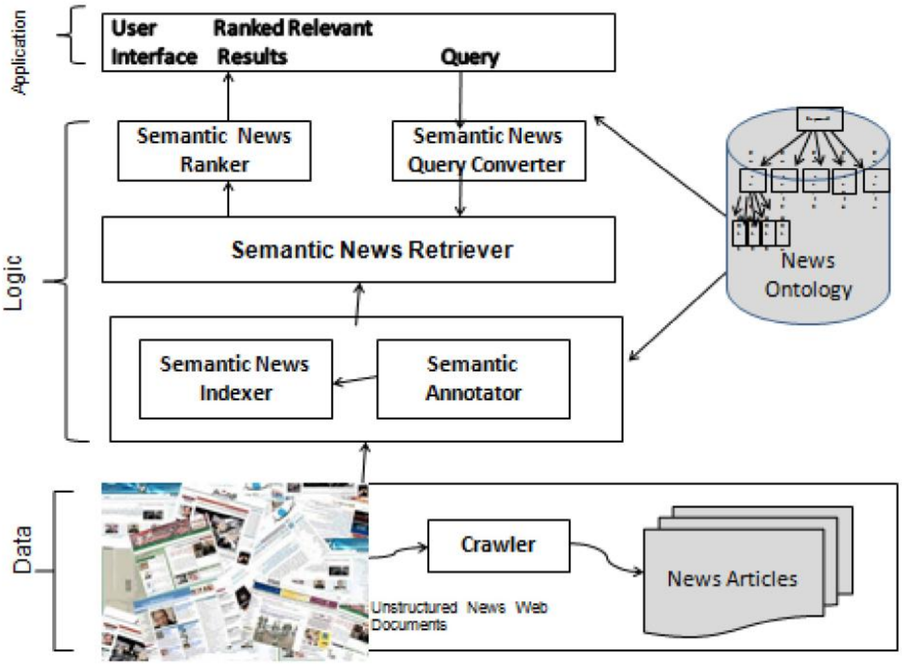


Figure 2: SNF Architecture

3.1 Data Layer

The Data Layer collects information from News websites with the help of a crawler program and stores in a web database for future retrieval. News agencies and other news related institutions or organizations set up web sites to publish their material for public and which is constantly increasing and therefore it is a suitable domain for semantic search.

3.2 Logic Layer

The logic layer incorporates four main components as Semantic Query Converter, Semantic News Annotator and Indexer, Semantic News Retriever and News Ranker. Each of these components are explained in the following.

3.2.1 Semantic Query Converter

The raw keywords entered by the user is preprocessed _rst and then expanded with the help of suggestions provided by the News ontology. The suggestions are given based on the concepts of the ontology or the links of the URLs or synonyms from thesaurus. The user can construct their query by making use of the suggestions provided. The pseudocode for the query expansion is given below.

```
Algorithm : Semantic Query Converter
Input : Raw Query Terms Q, News Ontology, News Knowledge base
Output : Semantic Query

Remove the special characters
Q = Q - Special characters
Remove Stopwords
Q = Q - Stopwords
Stemming
Q = Porterstemmer:Stem(Q)
Expand the query using
i) Suggestions from the Concepts of the Ontology
Q = Q U Ontology Concepts
ii) Suggestions from the web
```

```
Q = Q U Web Suggestions
ii) Synonyms from thesaurus
Q = Q U thesauruses : Synonyms(Q)
iii) Links of the Websites
Q = Q U extractlinks(URL)
```

3.2.2 Semantic News Annotator and Indexer

The NewsWeb database created by the data layer is preprocessed with i) HTML Parser which removes the meaningless HTML tags ii) Stop word Remover that removes less meaningful stop words iii) Stemmer to reduce the derived words into their stem. Preprocessor will also remove the advertisements, media (audio/video) contents etc.

The cleaned News Web documents are mapped with their semantic entities (concepts/meanings) of predefined News Ontology. Those semantic entities are also the connectivity to the Web News database rather than the keywords alone. After analysing large number of News articles of news websites, the ontology is restricted with the following concepts.

- Astrology
- Business & Finance
- Crime,Law
- Disaster
- Editorial
- Education
- Environment
- Health
- Location
- Period
- Politics
- Regional
- Science & Technology
- Sports
- Weather

The subject of each article is indexed with one or more of the concepts above. The mapping score which indicates how good a web document is mapped to an ontological concept is computed. Standard language processing and information extraction techniques are used to annotate documents and extract ontology driven meta data files.

The indexer counts the actual number of occurrences of each topic and also finds the parent classes for each topic. The score is a function of term frequency and tag based frequency of keywords and concepts. The algorithm used to create the semantic base is herewith presented.

```
Algorithm : Semantic News Preprocessor and Annotator
Input : Set of News Websites, News Ontology
Output : News Indexed Knowledge Base

for each web document
Remove the HTML tags with HTML tagger
Remove Stopwords
D = D - Stopwords
for each word in D
stem = porterstemmer::stem(word)

/*Calculate the term frequency in the content*/
for each document
   for each keyword
      Tf(w)=count(w,D)

/* Calculate the term frequency in the specified tags */
Tagf(w)=count(w,Tags)

/* Annotate the contents with the News ontology */

i)Extract the keywords from the website
ii)The keywords are matched with one or more concepts of the ontology
```

```
iii)The keyword,concept pair is indexed with the websites

/*  Calculate the concept frequency in the content */

for each concept in the ontology
Tf(c)=count(c,D)

/* Calculate the concept frequency in the specified tags */
Tagf(c)=count(c,Tags)

/* Calculate the inverse document frequency */
idf = log Number of Documents / df
```

3.2.3 Semantic News Retriever

News agencies and other News related organizations publish their material for public in their websites which is continuously increasing. With the help of the News Ontology the Semantic News Retriever matches the News knowledge base and the extended Semantic News query. The search is performed on the meta data instead of the actual pages and a list of relevant URLs are returned. A number of domain specific rules as well as set of heuristics are used to enable semantic searching. The matching strategy procedure is written based on the hierarchy of classes in the News Ontology. The index structure used for the purpose of retrieval stores the Query words of each query and the matching Url set for each word which then produces the result as intersection of url set.

```
Algorithm : Semantic News Retriever
Input : Semantic Query, News Knowledge base
Output : Relevant Ranked Result

Hash Index table h(i,j) with two columns
for each word i in the semantic query
{
   Insert the query words in the first column of the index structure
    h(i).add = QWi
    Find the matching URLs from the knowledge base
    for each url j
    {
          Insert the URLs in the hashtable corresponding to the query word
          if   Qwi matches with the url
          h(j).add = h(j) U Url
    }
}
Calculate the term frequency of the query words
Find the intersection of all the URLs
result = h(1,2) for each word i in the hashtable
{
result = result ^ h(i, 2)
}
calculate score1 as tf (words) + tf (concepts) * idf * idf
calculate score2 as tagf(words)+tagf(concepts)
Present the results in the order of score1,score2
```

3.2.4 Semantic Ranker

The retrieved URLs of Semantic News Retriever is ranked with the relevancy of the user constructed query. During this comparison, the topics of the initial query are taken into account. This is because when searching for relevant pages, one should consider what the user intended to find in the first place. In all the comparisons the weights of topics are taken into account so as to determine the relevance between the pages. This relevance is measured with the weights calculated by the improved dynamic ranking algorithm. The weight is a function of term frequency and the tag based frequency. The term frequency is the local weighting factor which reacts the importance of the term within a particular document. The global weighting factor considers the importance of a term within the entire collection of documents knows as news frequency(nf). The inverse news frequency (inf) which relates the document frequency to the total number of News articles in the collection (N) is computed. As semantic rank algorithm these values are calculated for the keyword as well as for the concept of the keywords.

The overall annotation and retrieval process is given in Figure. 3. The weight of a term in a document is defined as a combination of tf the inf. Now the similarity coefficient (sc) between the query and the web document is defined by the dot product of weights of the words and semantic entities of the semantic query (t). and the web document.

$$Sc(Q,Di) = \sum_{j=1}^{t} wt(w)qj * wt(w)ij + \sum_{j=1}^{t} wt(c)qj * wt(c)ij \quad \text{….. 1}$$

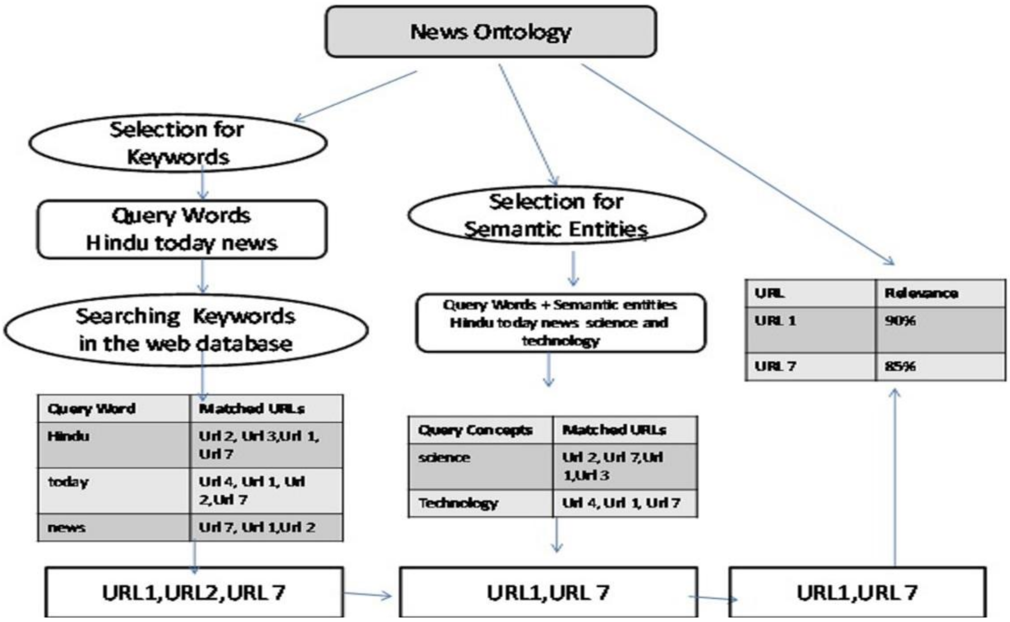
$$wtij(w) = tf(wij) * log\frac{N}{nf} \quad \text{…. 2}$$

Based on the similarity values the most relevant results identified by a threshold value are presented to the user. In the improved algorithm the weight depends on the two main factors. One is the quoted frequency of the keywords in the web documents and another one is the semantic entities in the content of the web pages. When there is no semantic entity the retrieval is nothing but the keyword based.

The improved algorithm based on TF-IDF algorithm can fit both traditional web and semantic web make the IR more accurate and promote the efficiency and the precision of traditional web search and semantic web search.

3.3 APPLICATION LAYER

It serves as the user interface for submitting the query and also used to deliver the ranked results to the user.


News Ontology
Selection for Keywords
Query Words Hindu today news
Searching Keywords in the web database
Selection for Semantic Entities
Query Words + Semantic entities Hindu today news science and technology
URL1,URL2,URL 7
URL1,URL 7
URL1,URL 7

Query Word	Matched URLs
Hindu	Url 2, Url 3,Url 1, Url 7
today	Url 4, Url 1, Url 2,Url 7
news	Url 7, Url 1,Url 2

Query Concepts	Matched URLs
science	Url 2, Url 7,Url 1,Url 3
Technology	Url 4, Url 1, Url 7

URL	Relevance
URL 1	90%
URL 7	85%

Figure 3: Web Documents Annotation with News Ontology

4. EXPERIMENTS

In order to collect daily articles, the RSS Feeds facility provided by the News Agencies websites are used. This results in approximately 70 articles collected per day under 10 different news sections. Approximately 5000 articles were imported in the database of the system to perform the evolution. Fig. 4 shows the distribution of the articles into the subjects. HTML parser is used to extract the plain text of the article by giving the corresponding URL as input. This dataset is submitted to the semantic annotation mechanism Table 1 shows a summary of the dataset and the annotations generated per news item. For the purpose of evolution 300 queries were posted on the system. The queries fell into four categories. 1) articles which refer to specific persons 2) articles about a specified location 3) articles containing one or more topics from the ontology and 4) articles containing topics of specified values for their attributes.

The results were evaluated manually with direct examination of the database contents and straightforward comparison with the complete article, when needed. The precision and recall measures were calculated for two phases. The initial search results were evaluated in the first phase and the relevant page results in the second one.

The comparison of keyword based and Semantic News Finder in terms of precision vs Recall is given in Fig. 5. This graph shows the general inverse relationship between precision and recall remains for both systems and the Semantic based system is with high recall and precision.

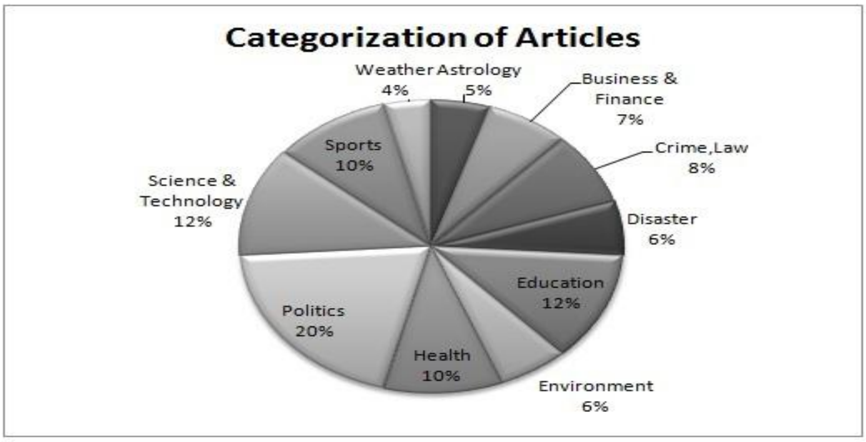


Figure 4: Categorization of Articles

Section	#news Items	#annotations	#annotations/item
Astrology	250	2400	10
Business & Finance	350	2578	7
Crime, Law	400	2845	7
Disaster	300	2520	8
Education	600	3542	6
Environment	300	2614	9
Health	500	3200	6
Politics	1000	12254	12
Science & Technology	600	4524	8
Sprots	500	3245	6
Weather	200	2654	13

Table 1: Average number of annotations per news item

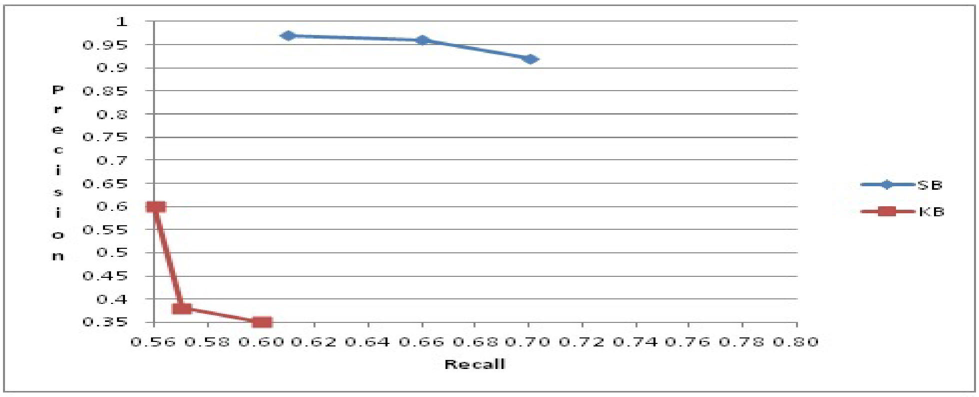


Figure 5: Precision Vs Recall

Fig. 6 a) and b) show the difference between the average Precision and recall values of 11 Queries, Mean average precision and Mean Average Recall values for Keyword based and the SNF respectively. The graph depicts that the SNF shows high precision and Recall compared to Keyword based. From this graph it is inferred that the proposed SNF outperforms the Keyword based in terms of precision and recall.

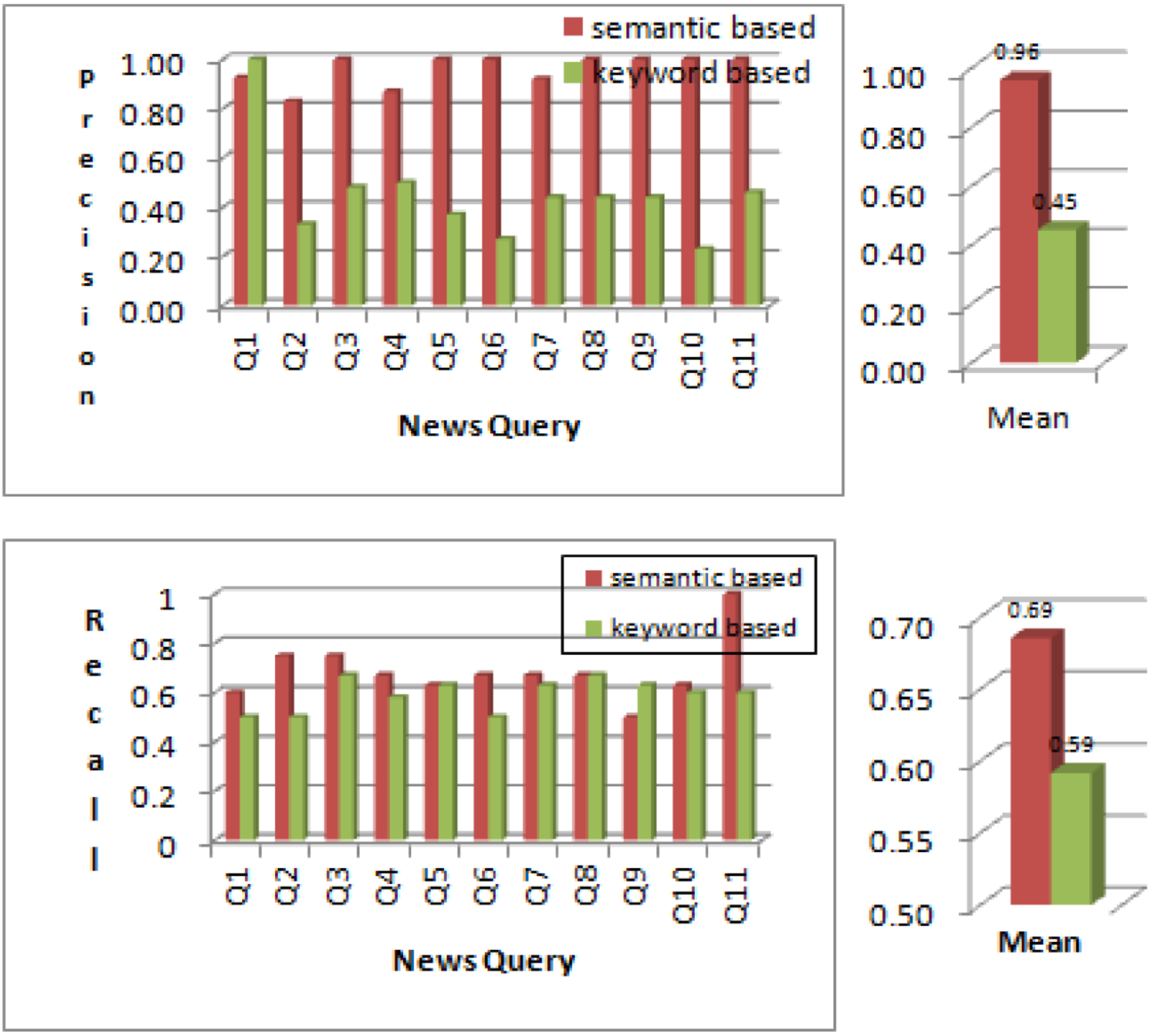


Figure 6: a)Result of Query matching-Precision Rate b)Result of Query Matching-Recall Rate

5. CONCLUSION

A novel semantic retrieval framework and its application in the News domain,which includes all the aspects of Semantic Web, namely Ontology development, information extraction, ontology population, inferencing, semantic rules, semantic indexing and retrieval. When these technologies are combined with the comfort of keyword-based search interface, a user-friendly, high performance and scalable semantic retrieval system is obtained. The evaluation results show that our approach can easily outperform both the traditional approach and the query expansion methods. The system can answer complex semantic queries without requiring formal queries such as SPARQL. Finally it is showed that how the structural ambiguities can be resolved easily using semantic indexing.

The current implementation can be extended and improved in many ways. First of all, the knowledge base is planned to enrich to support multiple languages. The performance will be further improved by implementing a word disambiguation module for lexical ambiguities. Finally a mechanism that expands the index automatically according to the user feedback is one of our future goals. The search engines available on the web do not resolve the problem due to the noise and the silence of web pages. The best solution is to exploit the documents semantic content by using ontologies. The solution proposed here is within the context; it uses the ontology in the indexing and research engine in order to enhance the research results relevance.

6. REFERENCES

[1] Ahmad Kayed, Eyas El-Qawasmeh, Zakariya Qawaqneh, 2010, RankingWeb sites using domain ontology concepts, Information & Management, 47 (2010) 350 -355.

[2] S. Chakrabarti, M. van den Berg, B. Dom, 1999, Focused crawling: a new approach to topic-specifc Web resource discovery, in: Proceedings of the Eighth International Conference on World Wide Web, Toronto, Canada, 1999,pp. 16231640.

[3] M. Diligenti, F.M. Coetzee, S. Lawrence, C.L. Giles, M. Gori, 2000, Focused crawling using context graphs, in: Proceedings of the 26th International Conference on Very Large Databases VLDB, 2000.

[4] R. Guha, R. McCool, E. Miller, 2003, Semantic search, in: Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary, 2003, pp. 700-709.

[5] H. Halpin, V. Lavrenko, 2009, Relevance feedback between hypertext and semantic search, in: Proceedings of theWorkshop on Semantic Search (SemSearch 2009), Madrid, Spain, April 21, 2009.

[6] J. Hendler, 2010, Web 3.0: the dawn of semantic search, Computer 43 (1) (2010) 77-80.

[7] Ivan Cantador, Alejandro Bellogin, Pablo Castells, 2008, News@hand:A semantic web approach to Recommending News, 5th International Conference on Adaptive Hypermedia, Springer verlag lecture notes in computer science, Hannover, Germany, July 2008, pp. 279-283.

[8] J. Johnson, K. Tsioutsiouliklis, C. Lee Giles, 2003, Evolving Strategies for Focused Web Crawling, in: Proceedings of 12th International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[9] Leonidas Kallipolitis, Vassilis Karpis, Isambo Karali, 2012, Semantic search in the World News domain using automatically extracted metadata files, Knowledge-Based Systems 27 (2012) 38-50.

[10] B. Masand, G. Lino_, D. Waltz, 1992, Classifying news stories using memory based reasoning, in: Proceedings of the 15th Annual international ACM SIGIR Conference on Research and Development in information Retrieval,Copenhagen, Denmark, 1992.

[11] C. Rocha, D. Schwabe, M. Poggi de Aragao, 2004, A hybrid approach for searching in the semantic web, in: Proceedings of International WWW Conference, New York, 2004, pp. 374-383.

[12] SYNC3 project, <http://www.sync3.eu>.

[13] M.Thangaraj, G.Sujatha, 2011, A Study on Searching Mechanisms in Semantic Web, in: International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004) 34, Volume 2, Issue 1, February 2011.

[14] M.Thangaraj, G.Sujatha, 2014, An architectural design for effective information retrieval in semantic web, in: Expert Systems with Applications 41(2014) 8225-8233.

[15] L. Zapf, N. Fernndez-Garc a, L. Snchez-Fernndez, 2005, The NEWS Project Semantic Web Technologies for the News Domain, in: 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, London, UK, 2005.

[16] Sharma, Manik, et al. "Design and comparative analysis of DSS queries in distributed environment." 2013 International Computer Science and Engineering Conference (ICSEC). IEEE, 2013.

[17] Manik Sharma, Gurvinder Singh, Rajinder Singh , Gurdev Singh. Stochastic Analysis of DSS Queries for a Distributed Database Design. International Journal of Computer Applications (IJCA) 2013.; 83 - 5:36-42