Query Architecture Expansion in Web Using Fuzzy Multi Domain Ontology

Shaghayegh Rabiee Kenari[a,*], Eslam Nazemi[b]

[a]Department of Computer Engineering, Qazvin Branch ,Islamic Azad University, Qazvin, Iran
Electrical and Computer Engineering Faculty, Shahid Beheshti University, Tehran, Iran

Abstract

Due to the increasing web, there are many challenges to establish a general framework for data mining and retrieving structured data from the Web. Creating an ontology is a step towards solving this problem. The ontology raises the main entity and the concept of any data in data mining. In this paper, we tried to propose a method for applying the "meaning" of the search system, But the problem for these methods is building a knowledge base that can be used for semantic search. The previous work interprets the query in three ways:'semantic relation in ontology', 'co-occurrence in the document', and 'semantic relation from Thesaurus'. The proposed method has two parts. The first part, using domain ontology for classified web pages based on keyword and the concept in each domain and builds Fuzzy ontology as Knowledge Base and the next section offers a method for expanding the query using built fuzzy ontology. In this paper, we tried to create knowledge base with WordNet as a comprehensive dictionary and extracted Sub string (phrases include multi words) from WordNet for each keyword in each domain ontology. The created Search engine was applied to an experimental system to evaluate the "precision – Recall" and it was revealed that applying the proposed method can improve query expansion 11%better in our experiments for precision.

Keywords: Semantic Search, Ontology, Query Expansion, Fuzzy Ontology

1. Introduction

Nowadays there is a vast amount of human knowledge in the form of electronic information and it is very difficult to find the desired information. One of the most important needs in today's digital world, information retrieval is discussed. IR refers to a process in which the user enters there quired information to retrieve the information that is relevant to their information needs. The growing of information on documents and text materials have caused more accurate and efficient case retrieval in recent years. Therefore, various methods have been proposed for optimal retrieval. The various models of retrieval aim to improve their call and precision.

As noted above, achieving more precise information on the web in accordance with user requirements is one of the most important challenges. Among the problems found in the search engines, which are the primary means for information retrieval and web mining, include:

- Some search engines are relying on only keywords to search.
- Inability to understand the relationships between words.
- Most of the retried data are not matched with the text is configured with the user's query.

These problems motivated researchers to help people by following two different strategies [1]:

- Changing the infrastructure of the current web to the semantic web.
- Placing the keyword based search engines as the base and doing some modifications to make them considering the query and web page context in order to improve their efficiency.

There was a big problem over the realization of the first idea. The problem was that there were already millions of millions documents in current web that should apply considerable modifications in their structure to express their content in RDF and RDFS [7]. On the other hand, for solving the problem of word sense ambiguity (one word corresponding to several different meanings and vice versa) and making a common understanding in a specified domain, diverse domain ontologies should be developed to cover existing documents in www. It is why that Quiz RDF combines traditional keyword querying of WWW resources

* Corresponding author. Email: rabiee.sh65@gmail.com

with the ability to browse and query against RDF annotations of those resources [2]. Search engines that are following the second strategy use keyword based search engines as their underlying layer and then add additional components to enhance their recall and precision [3,4,5,8]. Our proposed architecture in this paper also follows the second strategy.

In the world around us, words have different concepts and different forms, Because of the nature polymorphic of words, each word can has different meanings and concepts in different domains of knowledge. So it is very difficult to understand the human request from machine. Much work has been done in order to understand human speech by machine. One of these methods will help to solve these challenges is Conceptual structure. Conceptual structure utilizes various concepts to understand the user's query. Also having different concepts and models to interpret the words in question deals. One of concept can we noted, is ontology.

In this study, we use domain ontology to better understand the concepts of the human world and their relations, do their tasks more precisely and separate pages. The proposed query expansion sub system helps to refine the queries. This subsystem has two major differences with existing query refinement components in other architectures:

A. Gathering basic information:

The query refinement component needs some source of information to propose new terms in order to refine a query. In our proposed architecture we use multi domain ontology to gather the basic information, in other words, on the basis of the concepts that are already defined in the domain ontology; the query refinement component looks for the web pages that are related to these concepts and stores them in a database for further processing; as a result, in spite of other existing query refinement components, there is no need to violate the user privacy through monitoring his behaviour or his files to know about his preferences. Furthermore, users do not need to take their time to fill out the forms to introduce themselves and their preferences.

B. Selecting appropriate terms from created data base to refine queries:

In existing architectures, query refinement components have to interact with users for selecting appropriate terms from database and adding them into a query; whereas in our proposed architecture, the proposed fuzzy ontology constructor subsystem calculates the fuzzy relations between terms which are extracted from the stored web pages in previous step automatically and then suggests the terms with highest membership degree to refine the query of each Search Agent.

The rest of paper is organized as follows: section II reviews the related works briefly. Section III will describe the proposed architecture in detail. Section IV elaborates how to construct the fuzzy ontology which is used in query expansion subsystem. Experimental results are found in section V. Section VI concludes the paper and presents future works.

2. Related Works

In the real world, because words can have many meanings, there are some approach relies on the concepts. These approaches can be divided into two categories.

- Search engines that use ontology. These approach use ontology for semantic interpretation of the user's documents and queries, they have inference engine. Some examples of this type of engine Watson [10], Sem Search [11] and Falcon [12].

Approach of each these engines briefly stated as follows:

a) Interprets the user query and extract relevant concepts

b) Extracted concepts are used to construct a new query. At this stage, we formulate the concept of the entered query.

c) Ontology runs on the user's query and then results are displayed.

- Another set of search engines, based on keyword-based search engine then integrate with ontologies on the higher layer. Top layer can include :

a)Domain-specific ontology

According to literature [25] PASS search engine uses a fuzzy ontology to help users to refine their queries and getting more relevant results trough the keyword based search engine. The ontology is built automatically and determines the fuzzy relation between the terms. In PASS architecture the fuzzy ontology is constructed from a collection of documents which are not collected based on domain ontology so the flexibility of the system on changing the domain is reduced.

b)Agent intelligent

Master-Web and AGATHE are the two combined patterns in traduced in [19]

c) User Profile

According to literature [15] Query expansion terms are extracted automatically based on user behaviour. Click Streams and the user's history are analyzed. Also in [18] a probabilistic method for query expansion based on the user interest model which automatically created and updated, was presented.

d)Correcting mechanisms requests

- Feedback
- Neural Network
- WordNet

All of above model, make a new way to interpret the query. Therefore, explain a new concept in semantic search engine as named Query expansion. The main problem in Query expansion is find a best knowledge base for expanding. In [26, 27] use synonym from domain ontology and [28] use sense of keyword from ontology to expanding, but these method retrieved vast domain for each keyword and at the end make noise on result. Another works that use domain ontology that convert query to formal query without considering polysemy that leading reduce precision. Another method to create knowledge base using Thesaurus. WordNet is a Thesaurus that include Synonyms, Hypornyms and other relation between words. many works was provided in this field. For example in literature [14, 16, 17] they apply Synonyms, Hypernyms and Hyponyms for expanding the query. In 2010, [13] presented A method of query expansion based on semantic relations which extracted words from wordnet that have semantically related with query, Then choose words with more value. In [20] Gonzalo et al. use a manually disambiguated test collection of queries and documents derived from the SEMCOR semantic concordance. Their experiment covers three types of index spaces: original terms; word senses derived from manual disambiguation and finally WordNet synsets. According to Gonzalo, indexing with wordNet synsets improves information retrieval by more than 29%instead of word forms. First research in WordNet was conducted by Voorhees. This research has shown that short queries have better results than long queries. It is also use feedback for query expansion make better result for long queries. Voorhees [22] carried out experiments to exploit the semantics contained within WordNet with sense of keyword to improve retrieval effectiveness by indexing with word senses instead of word stems. The results showed that the effectiveness of the vectors produced by this disambiguation technique was worse than word stem vectors for all five collections. Navigli and Velardi[23] use sense information and ontologies for query expansion. They argue that expanding with synonyms and hyperonyms has a limited effect on web information retrieval performance. They suggest that other types of semantic information derivable from an ontology is more effective such as gloss words and common nodes. This is because words in the same semantic domain and same level of generality are best candidates for expansion.

The problem with domain-independent ontologies such as WordNet is that because they have a broad coverage, ambiguous terms within the ontology can be problematic. For narrower search tasks, domain-specific ontologies are the preferred choice. Domain-specific ontologies have been constructed in many different application areas such as law, medicine, archaeology, agriculture, geography, multimedia, business, economics, history, and even the news domain to name but a few. For example, Nilsson et al. [9] use a domain specific ontology based on Stockholm University Information System (SUiS) to carry out query expansion. SU is differs from other question answering systems because it does not allow free-form questions. The question types are restricted to who, what, when and where. Instead of expanding queries with all semantic relationships provided by an ontology such as WordNet, only synonyms and hyponyms are used to increase precision. The experiments have shown an improvement in results.

3.Proposed Architecture

This section is dedicated to elaborate our proposed architecture for domain specific search engines base fuzzy ontology. What makes this architectural model different from all other existing architectures is a query refinement component which helps Search Agents to refine their queries and express them in a more precise way while interacting with their underlying layer. This query expansion subsystem is applying fuzzy ontology to help the Search Agents. The following is an introduction to the architectural model's components as well as how they relate to each other and the workflow. Main stream Combination search engines work is composed as follows Fig. 1.

To increase accuracy, and optimize the results of user requirements in the proposed method, we add several parts to this process. Fig.2 shows the added items.

The process observed completely different part in fig. 3. In the following section we introduce the components of the proposed architecture, and then we describe how the components workflow.

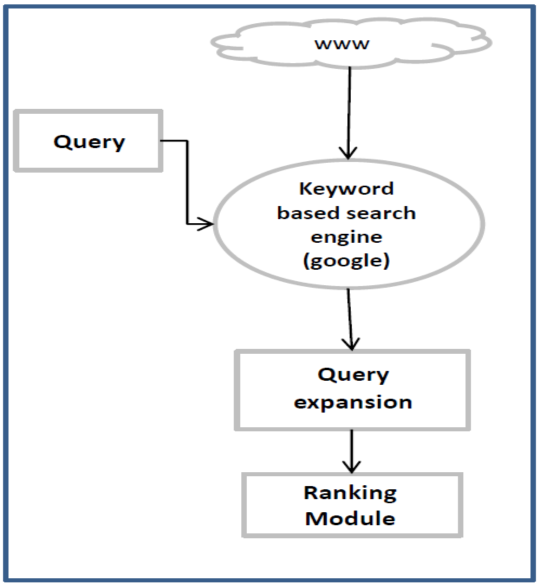


Fig. 1.The general architecture of a semantic search engine

A. Ontology

Ontology contains a huge collection of concepts and word. If we show it in a different domain, we have a wide range of words in each area. There are two ontologies in our proposed architecture. At first we use multi domain ontology that consist concepts and relation between the concepts. The domain ontology uses to assign the document to each domain. The second ontology is a fuzzy ontology which helps Search Agents to add more specific terms to their queries and each term in ontology with WordNet to get more relevant results while interacting with keyword based search engines.

B. Crawler

Crawler retrieves all pages from the web. This step occurs at Offline. After retrieving the pages and put them in a database, we create dictionary of pages. In general, to create a dictionary, usually using conventional algorithms such as porter for obtain the root of each words in each document. In this paper, because of using domain specific ontology and WordNet , we're required to use the original words in dictionary.

C. *Pages Separation*

Traditional search engines lack mechanisms for pages category. Hencethe pages can be separated by applying the methods to reduce the search space and it's improving the search process [21]. We compare each domain ontology with keywords in each page. Then, store every word that was similar with regard to its domain and calculate weight of domain with (1).

$$\omega = \frac{\sum TF}{n} \qquad (1)$$

N: total number of words in the document

TF : Repetition frequency words that the words were same with ontology

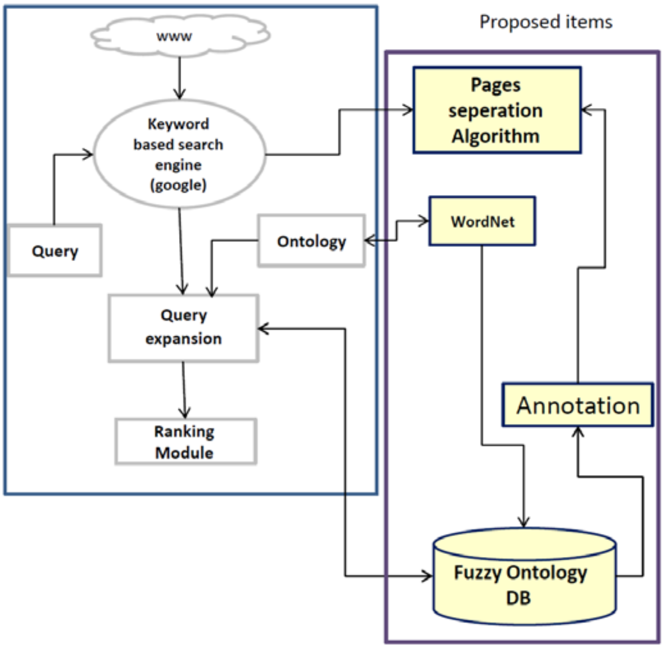


Fig. 2. Added items to the search engine

D. Preprocessing

In this step we annotate pages witch and ideates fuzzy ontology. In order to highlight a subject of a page we use annotation, it leads to access easy and faster to data.

E. Response to user

o *The user interface*

This interface enables the interaction between the users and the database which contains the classified web pages that are categorized based on the domain ontology. Users send their requests to the system through the user interface and get the related URLs.

o *Query Expansion*

When the user enter the query through a user interface, this query separate to multi keywords in multi domain. Then choose related substring from fuzzy ontology DB.

F. Ranking Module

In this part when a query entered, at first, eliminate the stop words, then use subtract between each keyword extracted from query, to determine default domain. At first find equal substring with query from Fuzzy DB. If there is substring same as query, choose this substring as first candidate term for expand, retrieved all pages descending that annotated with this Substring. Then try to find another substring with most degree for each keyword and select the default domain. Pages that contain these Substring should be belong default domain to be selected. There are some classes and their relationships in each domain ontology. Each class stands for a concept in the domain. For each class in the domain ontology, a series of related terms are defined.

As mentioned above, we want to classified pages with these domain, so after retrieve pages from web, in offline, we classified these pages. After classified, find substring for every word in each domain ontology via WordNet API. Then annotate every page with these substrings. So when user import query, at first, search engine should find keyword from query. When query were separated to some keyword, define domain of each keyword and extract their substring from fuzzy ontology DB and retrieve document that consist these substring.

4. The Proposed Fuzzy Ontology Constructor Subsystem

The purpose of this subsystem is arranging phrases in a hierarchical manner hence whenever a user defines the query; this subsystem offers a list of expressions with varying degrees. They also present in more pages and with more frequency than the keywords in the domain ontology. fuzzy ontology subsystem architecture is shown at Figure 4.Search keyword in domain ontology in WordNet. WordNet can be used for many different semantic relations. words with co-occurrence relationships are the output of this project from WordNet. We can retrieve these terms as Substring from WordNet. Each word in domain ontology is part of each Sub String. Words such as "hospital bed","hospitality "and " mental hospital"are Substrings that obtain for "Hospital" from WordNet. These terms when a query presented, offers to the user, then according these phrases, can select his intended phrase. This causes the user to be able to properly phrase their intended question or phrase to search for close to what it considers.

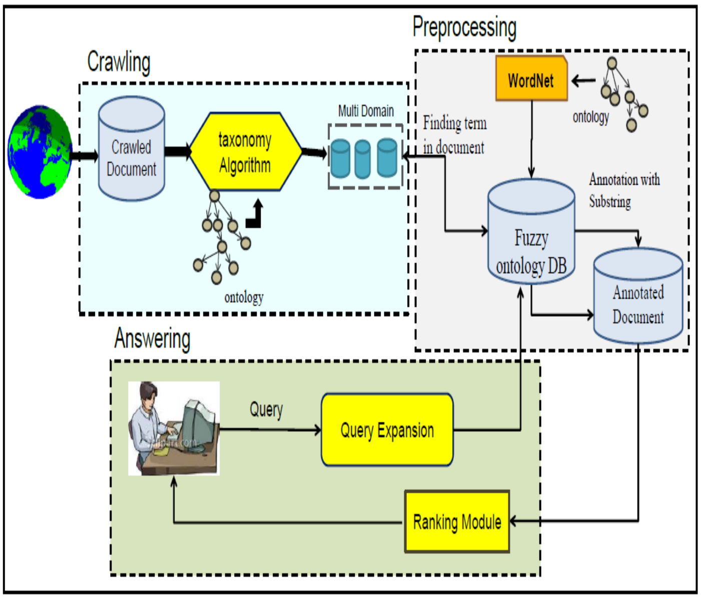


Fig. 3. The proposed architecture

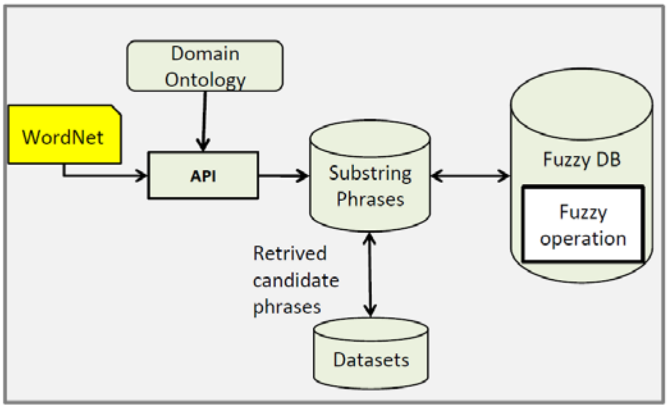


Fig.4. extract candidate terms and create fuzzy ontology DB

Suppose that $T= \{t_2, ……., t_n\}$ is a collection of keyword in each domain ontology. Obtain Sub String of these words through WordNet API and store them in a database. Then follow pages separately for each domain that contains collection of *T*in each domain. Suppose that $C=(u_1,u_2,...,u_n)$ is a collection of *URL* that each shows= (t_1, t_2, …, t_n) that is ontology 's keyword.

Search extracted Phrases from WordNet in any page that containing the keyword. If consider the, the frequency of occurrence an in *u* is displayed occure (t_j, u). Now, we must purify substring in the WordNet DB; Keep phrases that equivalent with phrases in document for fuzzy operations and pass up rest of them.

- *Description of Fuzzy operations:*

Let suggests SK (t_i, t_j) show that, cover more specific of interval than .Membership degree of SK (t_i, t_j) which shown as μ_{SK} (t_i, t_j), that is defined by (2).

$$\mu_{sk}(t_i,t_j)=\frac{\Sigma\mu_{occur}(t_i,u)\otimes\mu_{occur}(t_j,u)}{\Sigma\mu_{occur}(t_j,u)} \qquad (2)$$

Where denotes a fuzzy conjunction operator. According to (2), if the frequency of occurrence is greater, we can say with greater confidence level that has more degrees of satisfaction than. This selectivity is due to this reason that , we follow words that the most commonly used in desired domain. To make fuzzy ontology, at first, membership values are calculated for each pair of distinct words by (2).In order to select appropriate words, we need a measure. In this step, degree of importance is calculated. Due to the weight of selected words, we consider the following formula to calculate the degree of importance:

$$\text{deg}_i=\sum_{i=1}^{n}\mu_{sk} \qquad (3)$$

This formula is acquired sum of the importance weights assigned to the selected word. It consider as the importance of weight. According to the obtained weights, the candidate phrases arranged in ascending order according to their weight and stored in Fuzzy ontology DB.

5. Experimental Results

A. Implementation details

Our proposed architecture is deployed in C#. Two domain ontology (computer and medicine) created in protégé, and then transform classes of ontology to XML format to use for implementation. C# API used for retrieve Substring from WordNet. Information is stored in SQL Server 2000 relational database system.

B. Data collections

As mentioned above, two domain ontology and two data set used in this implementation. “ohsumed “dataset with 5400 document used for medicine domain and dataset that used for computer domain called “my data” with 1500 document.3402phrasesextractedfor two domain from WordNet .

C. Parameters used for evaluation

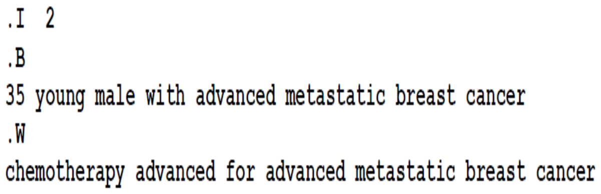
Precision and Recall are the two parameters mostly used for evaluating the efficiency of search engines; where Precision can be seen as a measure of exactness and Recall is a measure of completeness. Often, there is an inverse relationship between Precision and Recall, where it is possible to increase one at the cost of reducing the other therefore Precision and Recall scores are not discussed in isolation. Instead, both are combined into a single measure, such as F-measure, which is weighted harmonic mean of precision and recall. Precision, Recall and F-measure are commonly evaluated as shown in (5), (6) and (7) respectively.

$$\text{Precision}=\frac{\{relevan+docs\}\cap\{retrieveddocs\}}{\{retrieveddocs\}} \qquad (4)$$

$$\text{Recall} = \frac{\{relevantdocs\} \cap \{retrieveddocs\}}{\{relevantdocs\}} \quad (5)$$

$$F_{\beta} = (1+\beta^{2}).\frac{precision.recall}{\beta^{2}\,precision + recall} \quad (6)$$

For evaluating our system, we used "original query terms" that include tree part. Figure 5 shows an example of this query. As shown in Fig5 each query consists of three parts (I, B, W). "I" represent the number of query, "B" original query and "W" more details about the request.



```
.I  2
.B
35 young male with advanced metastatic breast cancer
.W
chemotherapy advanced for advanced metastatic breast cancer
```

Fig.5.one example of query

Since the demand simulation was performed to evaluate the improved the overall system performance after adding the component are requested to Improve; it is therefore necessary to run the system implemented in the following three modes according to the results of the evaluation parameters of the system to calculate and compare:

- System respond to questions without using an ontology.
- System responding to questions with query expansion using ontology(PSSE)[24]
- System respond to query with using fuzzy ontology and Co- occurrence relations[25].

For evaluating our proposed architectural model, we launched developed system using 12 search queries. Then we calculated Precision and Recall for each query with using equations (5) and (6).Figure 6 and Figure 7 shows a comparison between the value of evaluation parameters of our proposed architecture and three other search engine that mentioned above. As noted above, F-measure is an other way to get evaluation the results. Figure8 shows a comparison between the value of F-measure. We can see that the query expansion component helps Search engine to do their task more precisely which improves the overall performance of system.

6. Conclusions

The goal of information retrieval systems, providing a model, that retrieve information closer to the user request .A rich knowledge base, can help to achieve this goal. Our proposed architecture, is a multi-domain model search engine, that uses Wordnet to provide knowledge base. This architecture uses domain ontology to specify which domain is supported by the search engine.

As mentioned in previous sections, this architecture uses key word-based search engines as its underlying layer, so for popping-up more precise results from the keyword-based search engines to upper layers, we used Fuzzy ontology to expand their query contexts. for finding the term to expand we use Substrings of wordnet, and use Fuzzy Operations for finding the best candidate of these terms. Simulation results show that Precision and Recall are actually very good.

The present work and its prototype can be extended with agent to expedite the process of retrieval system. Using richer ontology can achieve more accurate result for our system. Also using semantic relationships between terms in the ontology for page annotations and indexing of the keyword leads to improve performance and can be considered as the future work.

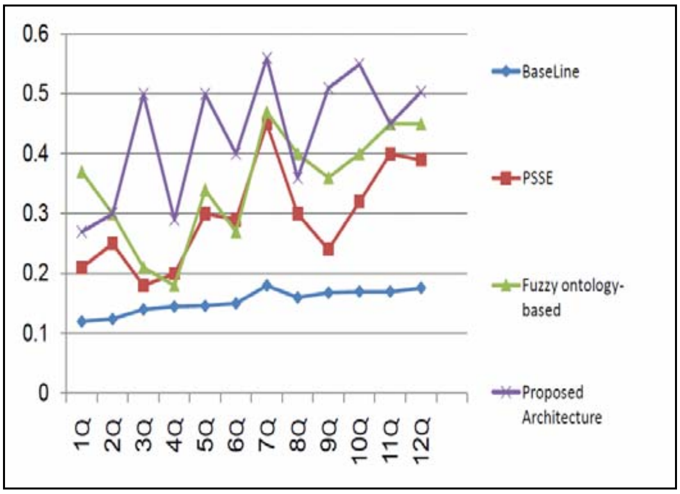


Fig. 6.A comparison between the value of evaluation precision of our proposed architecture and other methods.

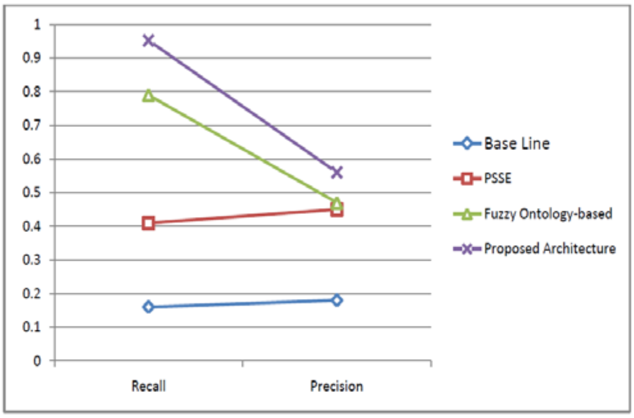


Fig.7. A comparison between the value of Recall- Precision of our proposed architecture and other methods.

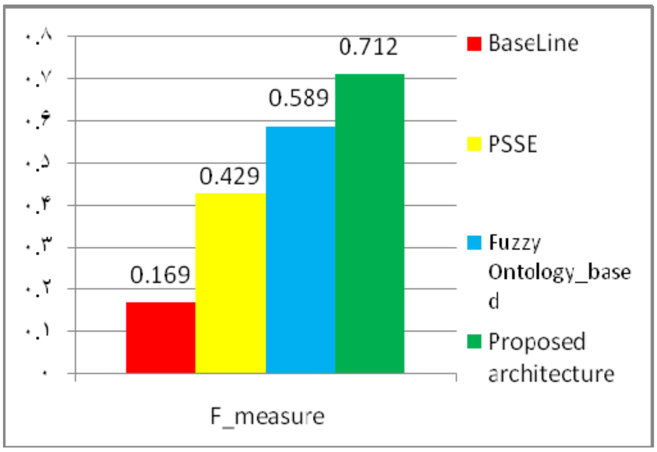


Fig.8.A comparison between the value of F-measure for our proposed architecture and other methods.

References

[1] T. Ramakrishna1,L. Gowdar, S. Havanur and B. Mllikarjuna, *"Web Mining: Key Accomplishments, Applications and Future Directions"*, In Proceedings of International Conference on Data Storage and Data Engineering, pp. 187 -191 , IEEE 2010.

[2] P.KOLARIandA. JOSHI," *Webmining:Research and practice*", Copublished by the IEEE CS and the AIP,2004,pp.49 -53, University of Maryland, Baltimore County, JULY/AUGUST 2004.

[3] J. Huang, J. XIU and J. Hong Gan," *THE RESEARCH AND IMPLEMENTATION OF WEB Spider IN SEARCH ENGINE"* ,pp. 244 – 247 , 2010 IEEE.

[4] J. Kassim,M.Rahmany, 2009, *"Introduction to Semantic Search Engine"*,Electrical Engineering and Informatics(ICEEI '09), 380 - 386 pp.

[5] M. C. Daconta, L. J. Obrst, K. Smith, *"The Semantic Web: A Guide to theFuture of XML, Web Services, and Knowledge Management."* ,John Wiley & Sons, 2006 .

[6] H. Haav, T. Lubi , " *A Survey of Concept-based InformationRetrieval Toolson the Web*" , Estonian Research Foundation *, 2001.*

[7] R.Kannan, 2010, "*Topic Map: An Ontology Framework for Information Retrieval*", National Conference on Advances in Knowledge Management, Volume 2, Page 195.

[8] J. Heflin, J. A. Hendler, S. Luke*." SHOE: A blueprint for the Semantic Web*". In D. Fensel, W. Wahlster, and H. Lieberman, editors, Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, pp. 29–63. MIT Press, 2003.

[9] K. Nilsson, et al. (2005). *"SUiS - cross-language ontology-driven information retrieval in a restricted domain.*", In Proceedings of the 15th ,NODALIDA conference.

[10] J.Euzenat, INRIA Grenoble Rhône-Alpes, France,*" Watson, more than a Semantic Web search engine"* , Semantic Web 2 (2011), pp.55–63 , DOI 10.3233/SW-2011-0031 ,IOS Press.

[11] Y.Lei ,V.Uren , and E.Motta ,*" SemSearch: A Search Engine for the Semantic Web"* , Knowledge Media Institute (Kmi), The Open University, Milton Keynes,pp. 1- 16.

[12] G. Cheng ,Ge, and Y. Qu.*" Falcons: Searching and browsing entities on the Semantic Web".* In Proc. WWW-2008, pp. 1101–1102. ACM Press, 2008.

[13] M. Shabanzadeh,N.Nematbakhsh,*"A Semantic Based Query Expansion to Search"*, International Conference on Intelligent Control and Information Processing , Dalian, China, pp. 523-528,2010.

[14] A. Yokoyamaand, V. Klyuev,*"Search Engine Query Expansion using Japanese WordNet"*, International Conference on Humans and Computers(HC2009),vol. 12, 2009.

[15] Chen ,Y.Du and Q. Peng, "*Extracting query expansion terms based on user 's search behavior*", Second International Symposium on Computational Intelligence and Design 2009.

[16] F.A. Grootjen, T.P Weide., , "*Conceptual query expansion*", Data & Knowledge Engineering V. 56,174-193, 2006.

[17] Y.Qiu, H.Frei-P. 1993, "*Concept based query expansion*", Proceedings f the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, Pittsburgh, Pennsylvania, USA, pp. 160-169.

[18] Z. Jiang, Z.Yu, 2010, "*A New Technology of Query Expansion Based on New User Interest Model*", International Conference on Future Computer and Communication,vol. 2 ,pp. 326-329.

[19] F.Freitas, L. Cabral, *"From MASTER-Web to AGATHE: the evolution of an architecture for manipulating information over the web using ontologies";* RECIIS – Elect. J. Commun. Inf. Innov. Health. Rio de Janeiro, v.2, n.1, p.73-84, Jan.-Jun., 2008.

[20] J. Gonzalo, et al. (1998)." *Indexing with WordNetsynsets can improve text retrieval"*, Coling-ACL 98.

[21] D. Mukhopadhyay ,A.Banik , S.Mukherjee , *" A Domain Specific Ontology Based Semantic WebSearch Engine ",* Advanced Communications & Networks Lab, Division of Electronics & Information Engineering Chonbuk National University 561-756 Jeonju, Republic of Korea.

[22] E. Voorhees, (1993).*"Using wordnet to disambiguate word senses for text retrieval".* ACM SIGIR, 171–180.

[23] R.Navigli, P.Velardi (2003*)." An analysis of ontology-based query expansion strategies workshop on adaptive text extraction and mining"* (ATEM 2003). In 14th European conference on machine learning (ECML 2003), September 22–26.

[24] S. Sabbeh , A. Riad , E. Hamdy, "*PSSE: An Architecture For A Personalized Semantic Search Engine*", International Journal of Intelligent Information Processing, Volume 2, Number 1, March 2011.

[25] D. H.Widyantoro, J. Yen, *" A fuzzy ontology-based abstract search engine and its user studies"*, Department of Computer Sciences Texas A&M University. IEEE (2001) , p. 1291-1294

[26] Y. Qiu., "*Concept based query expansion.*" . Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press. Pittsburgh, Pennsylvania, USA**:** , pp. 160-169, 1993.

[27] F.A Grootjen,*"Conceptual query expansion Data & Knowledge Engineering"* ,. V. 56**:** 174-193, 2006.

[28] R. Rahul,Y. A. Joshi, "*Concept-based Web Search using Domain Prediction and Parallel Query Expansion*". Information Reuse and Integration, IEEE International Conference,. Waikoloa Village,pp. 166 – 171,2006.