IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Semantic Indexing Techniques on Information Retrieval of Web Content

Abdullah K-K. A[1], Robert A.B.C[2], Adeyemo A.B[3]

Olabisi Onabanjo University, Ago Iwoye, Ogun State, Nigeria[1]

University of Ibadan, Ibadan Oyo State, Nigeria[2, 3]

Abstract: Information on the Web is increasing every day, the searching problem also increased. Indexing involves constructing a structured access to web content to facilitate search result. Retrieving relevant result is difficult due to traditional keyword based search which lack semantic. To address this issue, queries need to be disambiguated by considering the context (concept) using semantic search terms to index the search engine. However, user query may reflect multiple domains of interest which may lead to collection of semantically related terms. This paper discussed query terms with semantic search for retrieval of web content using different semantic indexing techniques. This would increase precision and recall for the semantic search when compared to traditional search. This would derive search related terms in the retrieved result taking into consideration the advantages, limitations on each of the techniques. Categories and Subject Descriptors [Information Search and Retrieval]: Clustering, information filtering, query formulation, Search process

Keywords: Semantic Search, WordNet, Ontology, Latent semantic indexing.

1.0 INTRODUCTION

Due to the huge amount of content and document repositories stored on the web, the problem of relevant search increases. The ability to access and retrieve relevant information remains a difficult task. The lack of efficient indexing method is still a major problem to information retrieval system. This is mainly because web is only understandable by human and the information cannot be processed by machine. The traditional retrieval systems have limited abilities to exploit the conceptualizations involved in user needs and content meanings due to inability to describe the relation between search terms. The search engines are keyword based which have not bridge the gap of vocabulary mismatch problem in retrieval system.

The word mismatch is a problem in the usage of natural language [1] User request need to be understandable by the retrieval system to avoid mismatch of terms, query may reflect multiple domain of interest. The indexers and the user do not always use the same terms, where synonymy terms may result in failure to retrieve relevant documents with a decrease in recall. Subsequently, polysemy may cause retrieval of irrelevant documents, thus decrease in precision retrieval. Subsequently, different researchers in the same field can name the same term differently [2]. This poses difficulty in text or large database expressing the same concept [3] The goal of semantic indexing is to use semantic information that is within the terms being indexed to improve the quality of information retrieval, unlike the traditional indexing method that is based on keyword matching. The use of semantic, index the concepts it contains rather than just the terms used to represent it.

Current information retrieval system needs to focus on using additional knowledge in order to retrieve relevant documents. The idea is that high-level of semantic content information can be accurately modelled using conceptual indexing so that related documents that may not share the same terms are still represented by nearby conceptual descriptors [4]. According to [5] described two categories of conceptual-based information retrieval approaches. The first category extracts semantic meaning from documents and queries by exploitingand model global usage patterns of terms. It forms latent relationship between texts terms and it known as Latent Semantic Indexing (LSI) [2]. This method is based on statistical approach and pattern that relies on co-occurrence information to enhance retrieval system such as weight loss, diet, exercise, eating right.

However, the second category employs the use of external semantic structures that map document representations to concepts. This is based on the concept of Semantic Web [6] which manually or automatically constructs taxonomy of semantic concepts and its relations for mapping documents and queries to taxonomy. The vital tools in searching for information and related resources in a Semantic Web (SW) is the Ontology. The ontology is used in information retrieval for query expansion, indexing and retrieval [7].

Different methods have been used to deal with vocabulary problem in query formulation such as interactive query refinement, relevance feedback, word sense disambiguation and search results clustering. But the most prominent technique is to expand the original query with suitable words that best capture the actual user intent.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

These knowledge would improve the semantic search, understand the intent of the searcher with contextual meaning of terms and generate more relevant results.
The rest of the paper is organized as follows. Section 2 discusses the related works on semantic indexing, document representation and similarity matrix. Section 3 describes different methods of indexing their advantages and limitation. In section 4, we conclude our work and recommendation.

2.0 RELATED WORK

Traditional techniques are unable to understand the meaning of the content, the documents and queries are represented by a bag of single words. This is due to the ambiguity, mismatch in the representation of the documents collection and imprecision in user queries which are inaccurate. However, semantic indexing is based on representation of the documents and requests by the senses of the words (concepts). This effect can be reduced in the retrieval system by utilising the conceptual indexing methods to improve the performance of a domain specific retrieval system. The approach raises the ambiguousness of the words and would solve the problem of disparity of the terms. In [8] proposed a semantic indexing model which exploits both the logical structures and the semantic contents of documents.
The semantic used in retrieval mapped words to concepts in its representation [9] where documents are represented as a set of concepts. Concepts usage will result in a retrieval model that is less dependent on the specific terms used however, [10] used the semantic knowledge instead of keyword-based index for efficient searching. Latent Semantic Indexing (LSI) exploits the use of statistical relation to determine the semantically relevant content as long as the relations are generated automatically from document. The degree of match between the query and documents retrieved are highly ranked in statistical approach [11] The assumption is that the more often terms co-occur in both the document and the query, the more relevant the document is with respect to the query [11]. However, the documents are represented as an index term based on the retrieval model to reflect its unique representation in a t-dimensional vector. Unfortunately, the method does not achieve greater performance. The semantic search aimed to solve the problem of keyword based model so that machine can understand the intent of the searcher. Semantic Web is a way to increase the precision of information retrieval systems. Therefore, additional external semantic structures (information resources) are needed for mapping document representations to concepts. Such resources can be dictionaries, thesauri and ontologies [12]. In {13],integration of knowledge through the use of WordNet to expand user' query over a collection with minimal textual information. However, [14] suggested that indexing with WordNet synsets can improve information retrieval. However, [15] presented a method of relating concept from ontology to the documents in the retrieval system. However, concept from one ontology would be different from concept in another ontology. Word sense Disambiguation techniques can use resources such as the Wordnet thesaurus [3] or co-occurrence data [16] to find possible senses of a word and map word occurrences to the correct sense.

3.0 SEMANTIC APPROACH FOR DOCUMENT REPRESENTATION

In this section, we describe semantic representation approach on document-concept matching. Concept based matching can be seenas an alternative approach to keyword matching. This can be used to tackle the problems of polysemy and synonym where both documents and queries are represented using semantic. The approach consists of building from a given semantic knowledge sources (LSI, Wordnet or Domain ontology), the concept pattern that best described the document content and the computation of term weights is influenced by the context. Representations of information resources and query are based on concepts so that words that are ambiguous can be disambiguated as in figure 1. The knowledge repository gives information about concepts and their relationships with other concepts. This will enable conceptual matching between extracted concepts that are relevant with the help of knowledge repository.

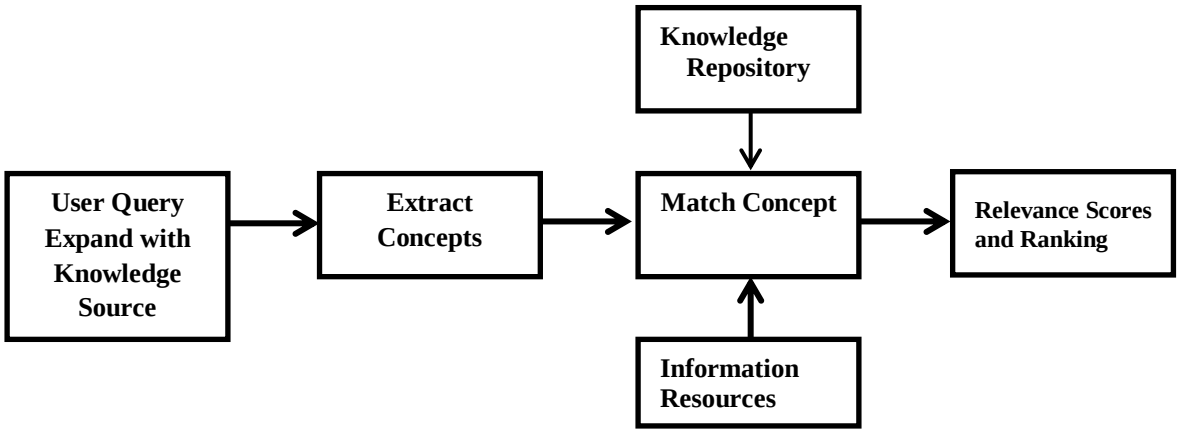


Figure 1: Conceptual Indexing Information Retrieval System

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

The set of terms in the documents can be grouped into a subspace but there is need to use an automated approach to achieve it. This can be achieved by clustering information according to concepts. Search results clustering is an efficient method to make the search results easier and it works better on snippets (a summary of the search results), which is different from document clustering [17].

3.1 Latent Semantic Indexing (LSI)

Using LSI to understand the semantic content of a document collection enables the definition of a logical semantic view. According to [18] described LSI to solve the problem of lexical matching using statistical derived conceptual indices. These are used to determine the set of terms, word relations (e.g. synonyms) and the strength of these relations. LSI is a similarity metric that is an alternative to word overlap measure and dimensionality reduction. But [19] presented a concept-driven algorithm for clustering search results, the Lingo algorithm, which uses LSI techniques to separate search results into meaningful group but do not consider building semantic relationships between the groups. However, common phrase are extracted using suffix tree clustering technique and concept induction using latent semantic analysis is conducted.

LSI can be represented in 2 or 3-dimensional space for visualisation, however, a mathematical technique named Singular Value Decomposition (SVD) is used for the representation. In LSI model, term-by-document matrix is performed by low rank and yields a new representation for each document in the collections. This represents projection into the latent semantic space, therefore, SVD decomposes an n- dimensional space (original space representation) into a k-dimensional (lower) space, this implies $n \succ k$. The SVD takes matrix A and represent it as $\hat{A}$ in a lower dimension space where A is orthogonal matrix U , a diagonal matrix Σ and transpose of an orthogonal matrix V :

$$A = U\Sigma V^T \qquad (1)$$

An orthogonal matrix U consists of term in each of the document collections and transpose of an orthogonal matrix V consists of document D in the new space, the diagonal matrix $\sum$ contains the singular values of A in descending order. Also, a user's query is represented as a vector in k-dimensional space and which are compared to document collection.

The matrix A can be represented as: $Term(t) = t_1, t_2 \cdots t_n$ that appear in each Document (d) = $d_1, d_2 \cdots d_n$ of a given query (q). The matrix A is decomposed so that U , V and $\sum$ can be found. Then, rank $(A) = r$ for $k_i \leq r$.

However, the new document vector and query coordinates in the reduced 2-dimensional space are found and finally the similarities (for instance Cosine similarity) between the rank documents in decreasing order of query are calculated.

This approach only works well with relatively small number of dimension with query compared to every document in the collection.

3.2 Wordnet Ontology

In order to have a better semantic, the knowledge resource (Wordnet) is used in generating context vectors. This knowledge resource understands the semantic content of the document and detects the concepts in ontology of the clustered documents. WordNet ontology is used to extract synonyms and hyponyms words and documents are allocated based on the concept of the ontology. Consequently, hypernym of each concept are detected and used to construct corpus related ontology. For each word or multiword concept candidate formed by combining adjacent words in text phrases, the ontology is checked using those words just as it is, however, the base forms are used which represents the document content better. This would enable reuse of resource although the corpus changes might omit some concepts with different form which might appeared in the source text and ontology.

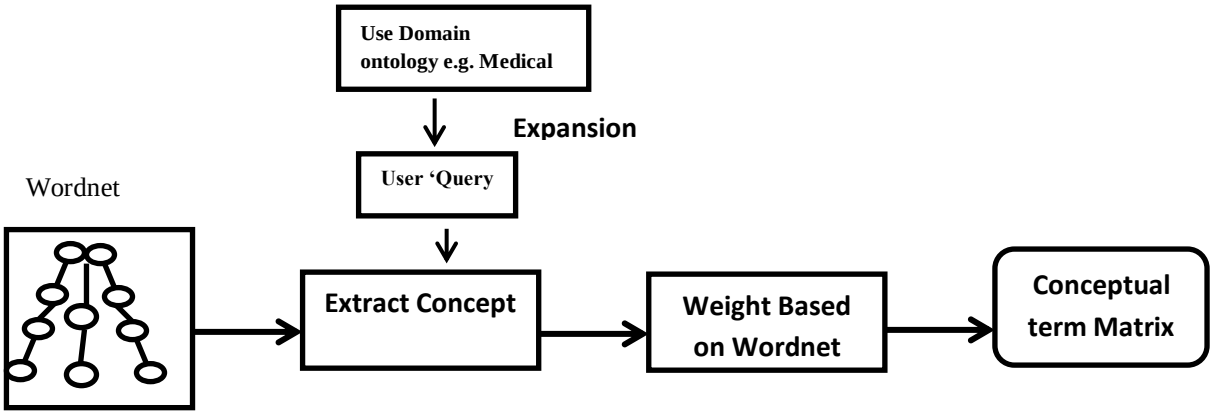


Figure 2: Concept-Based Term Weighting

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Using Wordnet approach introduce additional source of term importance which can be used for term weighting. This is used to find the conceptual information of each term using Wordnet ontology, however it determine the generality and specificity for a term such as the senses, synonyms, hypernym and the hyponyms (sub-concepts). Concepts from the Wordnet ontology are extracted from the user 'query as in figure 2 and identify those that occurred in the document collection, these are used as a substitute to idf. However the base form of each words are used such as rain, rains, raining etc. The Concept-Based Weighting value of a query determines the importance of a term.

3.3 Domain Ontology

Based on the Wordnet ontology, domain Ontology can be used to assist in the formulation of the user query and provide access to documents collection. From the figure 2, the approach expands the user'query by exploiting the rich semantics of ontologies [20]. Domain-based representation has been recently used in [21] which exploits the hierarchical IS-A relation among concepts that indicates the meanings of words based on the hierarchy. For instance, the hierarchy of "medicine" would have been indicated from the ontology hierarchy. This is done by mapping the query $c_q = (c_1, c_2 ... c_n)$ to the ontological user profile. Each query context is semantically related to concepts from the ontological user profile. News domain was used to expand the query by introducing several form of normalization for the semantic constituent with respect to the length of lexical chains and the size of the documents [22]. An experimental analysis was performed on queries that were refined using the domain independent ontology and query was given directly to the search engine [23], the experiment shows that domain ontology gave more accurate results than that of the direct search when structured.

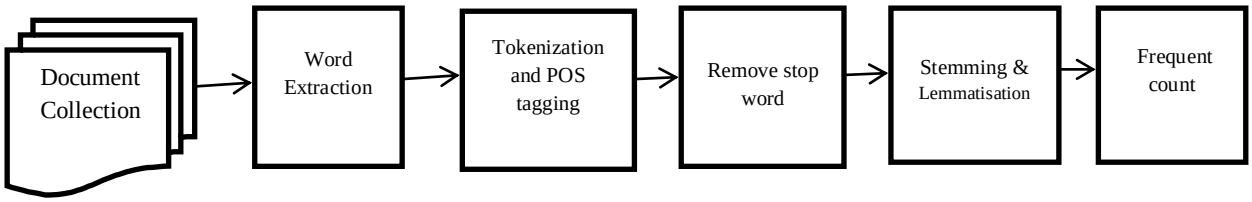
4.0 DOCUMENT INDEXING

Documents are retrieved based on the cluster the best related to the pattern or taxonomy, this will discard the irrelevant document in the collection. In the vector space model, document representation is based on query and content information. The integration of query and document space is the document-representation vector space model which improves the reliability and accuracy of the feature terms of documents. However, the documents representation (indexing) is one of the preprocessing techniques that are used to reduce the features complexity of the documents.

i. Feature Extraction

Documents can be represented by a wide range of different feature descriptions. There are two kinds of processes involved in text documents representation; these are document indexing and term weighting.The first step of the documents representation is to extract text document from the collection so that text preprocessing techniques are applied on it. The part-of-speech (POS) tagger processes the token and attaches a part of speech to each word. The tagger is used to tag the retrieved document to perform linguistic transformation using Natural Language Toolkit. A training set is created by tagging sentences with the correct senses of its contained words. The POS tagging gives information about semantic constituent of a word and the structure corresponds to meaning units and semantic analysis is performed.

Berry (2007) explained that information extraction willinvolve many complexities if tokenisation process is not done in the retrieval.



Word normalisation involves stemming and lemmatisation but these techniques produce a normalised form of web documents retrieved. . For example, the words "train", "training", "trainer" and "trains" can be replaced with "train". While lemmatization replaces the suffix of a word with a different one or removes the suffix of a word completely to get the basic word forms (lemma). A Wordnet-based which belongs to the group of dictionary lemmatisation algorithms is used.

ii. Feature Selection

Using semantic indexing techniques are used to identify the meaningful concepts. The documents are allocated to these concepts using cosine similarities. The techniques naturally create overlapping groups and well handle cross topic documents. In the document, vocabulary extraction process used concept frequency-inverse document frequency (cf-idf) for its representation. Concept (C_j) in the documents is represented as the concept frequency-document. The document (D_i) represents the document vector and each document selected is related to the cluster chosen $\{D_i = (w_{1i}, w_{2i} \cdots w_{ni})\}$ The weight (W_{ij}) is a function of the term frequency, collection frequency and normalization factors.Hence, the weight of each concept is calculated. The weight of concept j in the document i, T is represented in form of Matrix A $(CF_{ij} \times D_i)$.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

$$W_{ij} = cf_{ij} \times idf_i \quad (1)$$

$$idf = cf_{i,j} * \log\left(\frac{|D|}{df(j)+1}\right) \quad (2)$$

where cf is the total no of concept in the documents and idf is:

Table 1: Concept by Document Matrix

c/d	D_1	D_2	• • •	D_3
C_1	c_1d_1	c_1d_2	• • •	c_1d_n
C_2	c_2d_1	c_2d_2	• • •	c_2d_n
•	⬬	⬬	• • •	⬬
•	⬬	⬬		⬬
•	⬬	⬬		⬬
C_n	c_nd_1	c_nd_2	• • •	c_nd_n

4.1 Computing Similarity between Query and Concept
A similarity measures can represent the similarity between two documents, two queries or one document and one query. Similarityestimates the degree of similarity of a document d_j with concept i to a query q as the correlation between the vectors d_j and q. The vocabulary with a low frequency, the value is low compared to that with a high frequency,The correlation between query and concept in the document collection can be quantified by different similarity measures. Since, $w_{ij} > 0$ and $w_{i.q} > 0$ $Sim(q, d_j)$ varies from 0 to 1. A variety of similarity or distance measures have been proposed and widely applied but the mostly commonly used is cosine similarity measure.

The query and documents are represented as concept vectors, the similarity of query and documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors.

$$Sim(\vec{q}, \vec{d_j}) = \cos(q, d_j)$$
$$= \frac{\vec{q}, \vec{d_j}}{|\vec{q}| \times |\vec{d_j}|} = \frac{\sum_i w_{i,j} \times w_{q,i}}{\sqrt{\sum_i w_{i,j}^2} \times \sqrt{\sum_i w_{q,i}^2}} \quad (3)$$

The cosine is a normalized dot product, therefore the documents are ranked by decreasing the cosine values.

4.2 Algorithm: Conceptual Pattern Extraction
Input: Snippets D returned by the web Search engine for query concept C

Output: Conceptual patterns C & Frequency F.

Step 1: Read each snippet D, store it in database and perform data cleaning operation.
Step 2: for each snippet D do
if term is same as C then replace C by X
end
Step 3: for each snippet D do
if $X \in D$ then extract all conceptual patterns from D
end if
end for
Step 5: for each conceptual pattern do
Perform stemming operation.
end for
Step 6: Find the frequency of all repeated extracted patterns.
Step 7: Return Patterns C & Frequency F.
Step 8: Compute Similarity

5.0 CONCLUSION AND RECOMMENDATION

From the review works, it shows that keyword-based search method produce low recall and precision, high dimensionality when querying the information retrieval system. With semantic based indexing techniques, retrieval of information produced better precision and recall are achieved by using conceptual pattern.

Using structured queries allow expressing more precise user's needs, leading to more accurate results. This reduced polysemy and takes care of synonyms term in the documents using conceptual methods. In future research, the techniques can be enhanced to measure semantic similarity in automatic synonym extraction, query suggestion and semantic retrieval.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

REFERENCES

[1]. Bendersky, M.; Metzler, D.; and Croft, W. B. 2010. Learning concept importance using a weighted dependence model. In WSDM, 31–40.
[2]. Landauer, T. K. and Dumais, S. T. 1997. A solution to plato's problem: The latent semantic analysis theory of the aquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
[3]. Voorhees, E. M. 1994. Query expansion using lexical-semantic relations. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Springer, 61–69.
[4]. Beyer, K., Goldstein J., Ramakrishnan, R. and Shaft, U. 1999. When is `nearest neighbour' meaningful. In Proceedings of ICDT-1999, 217-235.
[5]. Egozi, O., Markovitch, S., and Gabrilovich, E. 2011. Concept-Based Information Retrieval Using Explicit Semantic Analysis. ACM Transactions on Information Systems. April 2.8: 1-34.
[6]. Berners-Lee, T., Hendler, J., Lassila, O. 2001. The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In Scientific American, Mai http://www.scientificamerican.com/2001/0501issue /0501 berners-lee.html. 285.5:34-43
[7]. Carpineto C.and Romano. G. 2012. A Survey of Automatic Query Expansion in Information Retrieval. ACM Comput. Surv.. DOI 10.1145/2071389.2071390 http://doi.acm.org/10.1145/2071389.2071390. January 44.1:1-50.
[8]. Chagheri, S. Roussey, C., Calabretto, S. and Dumoulin C. 2009. Semantic indexing of technical documentation, LIRIS 2009.
[9]. Haav, H. M., Lubi, T.-L 2001. A Survey of Concept-based Information Retrieval Tools on the Web. In Proc. of 5th East-European Conference ADBIS*2001, 2: 29-41.
[10]. Vallet, D., Fernandez, M. and Castells, P. 2005. An ontology-based information retrieval model. In The Semantic Web: Research and Applications, ESWC, 455–470.
[11]. Grossman, D. and Frieder, O. 2004. Information Retrieval: Algorithms and Heuristics," Second Edition; Springer Publishers, ISBN 1-4020-3003-7, 1-4020-3004-5, 2004.
[12]. Guarino, N., Masolo, C., and Vetere, G. 1999. OntoSeek : content-based access to the web". IEEE Intelligent Systems, 14:70-80.
[13]. Manuel, D. Maria, M. Alfonso, U. L. and Jose, P. 2010. Using WordNet in Multimedia Information Retrieval. CLEF 2009 Workshop, Part II, LNCS 6242, Springer-Verlag Berlin Heidelberg. 185–188.
[14]. Gonzalo, J., Verdejo, F., Chugur, I., Cigarrán, J. 1998. Indexing with Wordnet synsets can improve text retrieval, in Proc. the COLING/ACL '98 Workshop on Usage of WordNet for Natural Language Processing,
[15]. Khan, L., and Luo, F. 2002. Ontology Construction for Information Selection In Proc. of 14th IEEE International Conference on Tools with Artificial Intelligence, Washington DC, November. 122-127.
[16]. Schütze, H. and Pedersen, J.O. 1995. Information retrieval based on word senses. In: Proceedings of the 4th annual symposium on document analysis and information retrieval, 2.1:45-65.
[17]. Zamir, O. and Etzioni, O. 1999. Document Clustering: A Feasibility Demonstration.Proceedings of the 19th International ACM SIGIRConference on Research and Development of Information Retrieval, 46-54.
[18]. Deerwester S., Dumais, S. T. Landauer, T. K. Furnas, G. W. and Harshman, R. A. 1990. Indexing by Latent Semantic Analysis," Journal of American Society of Information Science, 41. 6: 391-407.
[19]. Osinski S.and Weiss D.. 2005. A concept-driven algorithm for clustering search results. 20.3:48– 54.
[20]. Ejiofor C. I, Williams E. E, Nwachukwu E.O, Weide T. 2013. Semantic Method for Query Expansion in an Intelligent Search System'' Journal of Emerging Trends in Computing and Information Sciences 4.6: 577-583.
[21]. Fernández M S. 2011. Semantically enhanced Information Retrieval: An ontology-based approach", Journal of Web Semantics: Science, Services and Agents on the World Wide Web. doi:10.1016/j.websem.2010.11.003. Phd.
[22]. Soni A., Sunhare H., Patel S. (2013):" Semantic Retrieval Technique Based On Domain Ontology " International Journal of Innovative Research in Science, Engineering and Technology 2 .1: 3187-3192.
[23]. Ruban S. and Behin S. S. (2015): An Experimental Analysis And Implementation Of Ontology Based Query Expansion . ARPN Journal of Engineering and Applied Sciences, 10.7: 3108 – 3111.