# Enhanced Semantic Search with Ontology-Based Information Retrieval of Web Documents

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

**[1]Olabisi Onabanjo University, Ago Iwoye, Ogun. [2]University of Ibadan, Ibadan Oyo State. [3]University of Ibadan, Ibadan Oyo State.**

## *Abstract*

***Retrieving desired information is still a huge challenge on the current web due to heterogeneity of information. The existing search engines returns a long list of results for the given query and are ranked by their relevance to the query. Ontology can be efficiently utilized for the bag-of-words that is often unsatisfactory as it ignores relationships between terms. In this paper, a framework is proposed on concept weighing for enhancing user query to improve search performance in order to satisfy the users need using domain ontology. It exploits conceptual hierarchy of the ontology with suffix tree clustering (STC) to construct feature vector for computing the similarity between weighted documents to improve the performance of the retrieved documents. The idea considers semantically related words from the ontology for increasing the quality of clusters. Experiments shows that there is increase in retrieval performance of high precision and high recall.***

**Keywords:**Hierarchical Ontology, STC, Feature vector, Clustering, Semantic Search.

## 1.0 Introduction

With the enormous amount of information present on the web, it has been difficult to find or access relevant information or documents. However, most search engines operate based on keywords of a query and web pages are considered as bag-of-words without perception of page creator's purpose and its concept. In response to the user's query, current available search engines return a ranked list of documents along with their partial content (snippets), this method limits the effectiveness of the search of relevant documents. The challenges can be dealt with in two ways: refine the user queries or optimize the search service so that the search service returns reasonable results whatever the user queries. Berners-Lee et al. [1] proposed the Semantic Web (extension of the WWW) in which information is tied to machine-readable information and automated services. Semantic web (SW) technologies pave way to the formulation of possible and effective solutions. Therefore, the most vital tools in searching for information and related resources in a SW are the ontology and intelligent agent. Ontology represents knowledge that could be understood by machine, used and shared among distributed applications and agents to improve knowledge management systems. Consequently, ontologies capture the semantic relationship between concepts or vocabulary used in a particular domain and can potentially be used to discover inherent relationships between descriptions of entities. Information Retrieval (IR) system returns a set of documents satisfying the information need expressed by user's query. The purpose of information retrieval is to retrieve all the relevant documents at the same time retrieving as few of the non-relevant as possible. Tradition text clustering methods use the bag-of-words (BOW) model for information retrieval [2] where single terms are used as features for representing the documents and are treated independently. Recently, researchers put their focus on the conceptual features extracted from text using ontology (Bag of Concept), the concepts are taken from the web page and represent the web page more semantically. But the concept features simply add related term to an entity by referring to ontology such as WordNet. These methods implicitly increase the dimensionality of the text data, similarity functions have been used to measure the diversity among the documents and between document and query [3]. Consequently, using WordNet ontology to exploit semantic relationship between terms can be substituted for terms with similar concepts which may be inaccurate in some concept due to ambiguity of concepts with loss of information.

In order to incorporate the semantic aspect in the search engine requires the use of ontologies entities (concepts and relation between concepts) as a search query and the retrieved documents can automatically group search results into thematic groups (cluster). Clustering techniques in information retrieval categorizes and organizes the search results into semantically

Corresponding author: Abdullah K-K. A.,E-mail:uwaizabdullah9@gmail.com,Tel.: +2348060046592

meaningful clusters in order to enhance the efficiency of retrieval system. It considers semantically related words such as synonyms, hyponyms or hypernyms for increasing the quality of clusters. Overall, the ontology can thus serve to increase either the exactness or the comprehensiveness of the matching process in IR systems which match representations of users' information needs to those of information objects. It is also recommended that post-processing of the resulting clusters be performed to eliminate groups with identical contents and to merge the overlapping ones. The feature-represented documents are transformed into a concept-represented thus achieve the processing of document clustering at the conceptual level. The conceptual features extracted from text using ontologies have shown that ontologies could improve the performance of text mining [4]. The concepts weight of each documents retrieved are estimated with references to the domain hierarchical ontology knowledge. Therefore, the feature vectors for each concepts of the ontology can be found.

The rest of the paper is organized into the following sections. The next section provides the background and related work. Section 3 explains clustering technique and section 4 discussed the proposed framework and finally section 5 discussed evaluation and conclusion respectively.

## 2.0 Background and Related Work

The Semantic Web uses ontologies as a structured representation of knowledge to improve information retrieval and to assist both humans and machines to better find information in web pages. One important issue on the management of documents and particularly the semantically supported document retrieval called the semantic search. Ontology-based semantic search rely on certain ontology structures (concept hierarchy). Concepts represented by ontology can usually be clearly depicted by a natural language because they both (ontology and the natural language) function similarly by describing the world. Most vocabularies used in ontologies are direct subsets of natural languages. Therefore, an ontological model can effectively disambiguate meanings of words from unstructured text, overcoming the problem faced in natural language where a word may have multiple meanings depending on the applicable context [5]. Similarly, it is necessary for a natural language processing system to be able to address syntactic and semantic aspects of natural language [6]. Hoang and Tjoa [7] surveyed several ontology based query systems on various aspects of using ontologies, including faceted search, query reformulation and refinement. Subsequently, to perform useful classification, the categorization is based on the actual information content or explicit representation of the information content of the source documents. The classification criteria must reflect the interest of the users. Clustering has proven to be an effective approach and useful technique that automatically organizes a collection with number of data objects into a much smaller number of coherent groups [8]. Researchers have applied standard clustering methods to web page clustering methods like hierarchical, partitioning, and many more. Modification of query is implemented as a recall-enhancing technique to extend the conventional bag of word (BOW) representation with relevant terms from WordNet. Experiments show the effectiveness of the extended representation with decrease in precision due to the possible noises introduced by the expanded terms. Kruse et al. [9] show that hypernyms are effective to be used as context words to indicate word senses for querying the Internet while [10] presents another method of indexing documents with ontology vocabulary based on the Vector Space Model, in this approach, a feature vector is calculated for each concept term from the ontology, based on the term's occurrences in the document corpus which is similar to our approach with modification in calculating feature vector. This way, the index terms derived from the ontology are adapted to the domain terminology. Tar & Nyaunt [11] articulates the unique requirements of text document clustering with the support of specific domain ontology using k-means so that the important of words of a cluster can be identified by the weight values. With the use of domain-specific ontology, the proposed system is able to categorize documents on the basis of the concept level. The method present a concept weighting that tries to capture some aspect of the Semantic Web. It is known that the performance of the k-means algorithm degrades with the improper selection of k values and initial parameters. Sureka et al [12] improved the Tar and Nyaunt [11] method by clustering web document using K-Means and DBScan algorithms, the performance showed that DBScan clustering algorithm using ontological weighting scheme produce better result than K-Means algorithm.

## 3.0 Suffix Tree Clustering (STC)

Clustering analysis is purely syntactical, it does not take advantage of the existing knowledge in the learning process and eventually, the most challenging problem is how the objects in a cluster should be clustered together. Clustering is thus preformed afterthe documents matching the query are identified. Consequently, the set of thematic categories is not fixed – they are created dynamically depending on the actual documents found in the results. The logic of text clustering algorithms is to use meaning of the words for the purpose of clustering thereby providing a basis for intuitive and informative navigation and browsing mechanisms. Words have meaning as an important property that relates them to other words, although there may not be a match of text string. Based on this idea, the suffix tree require the meaning of word strings not characters by keeping track of all n-gram of any length in a set of word strings. It allows strings to be inserted incrementally in time linear to the number of words in each string. The algorithm is fast with a runtime of *O(n)* and itoffers more semantic representation of the text present in the document. Suffix Tree Clustering (STC) is based on the Suffix Tree Document (STD) model which was proposed by [13], the algorithm used agglomerative hierarchical document clustering to perform the actual clustering.

The STC algorithm was used in their meta-searching engine to cluster the document snippets returned from other search engine in real time. The algorithm is based on identifying the phrases that are common to groups of documents. Accurate clustering algorithms require a precise definition of the closeness between a pair of objects, in terms of either the pair-wise similarity or distance.

## 4.0 Approach

### 4.1 Ontology-based Searching and Clustering

Ontology-based approach allows the representation of complex structure of entities that implement the knowledge about hierarchical structure of ontology as well as the information about relationships between entities. In this method, we refer to hierarchies and networks of concepts, we use both the terms link and edge to refer to the relationships between nodes**.** We enhanced our search queries by using domain ontology entities. These entities in the hierarchical domain ontology are used as queries on the search engine (Google) with suffix tree algorithm (Fig 1 entities), different document collections are automatically retrieved in response to the query (users' information needs). The documents retrieved in response to a query are sorted out according to the relevance of the query (most relevant), this enhance search result and extract useful knowledge from the domain. The documents retrieved (snippets) must be pre-process to remove unwanted terms and symbols (html tags), stemmed to remove duplicates of the same word or phrase by finding their common root (e.g. running, ran to become run) and a part of speech (POS) tagger is used to tag the retrieved documents (snippet) to perform linguistic transformation. However, most of the clustering algorithms aim to reduce high dimensionality while maintaining the document's semantic structure [14]. Therefore, using ontology-based for document clustering, suffix tree algorithm would behave better in a given domain (as discussed in section 3).

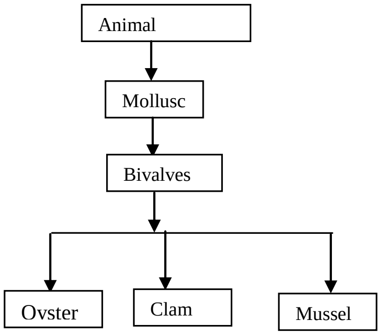

**Fig1. A small fragment of hierarchical structure of mollusc ontology**

For web search results, clustering the snippets can be obtained directly from a web search engine using Carrot2 API with 30 documents for effective clustering and avoid repetition of document (reduce high dimensionality). Semantic Suffix Tree clustering algorithm extracts the common word sequence from the document and uses common word meaning sequence to perform the compact representation by combining similar document together to form labels for each cluster as represented in Fig2. Therefore, the target document corpus will be clustered in accordance with the concept represented one and thus achieve the processing of document clustering at the conceptual level.

### 4.2 Document Modelling

The approach would be based on pragmatic use of ontology by relating the entities (domain semantics) with the actual terminology used in the hierarchical structure of ontology which represents the domains of interest, the idea is that this ontology can be used to disambiguate search. However, each label of the clusters is calculated for domain document relevance, multiple associations are allowed using semi-automatic method because relevant documents may contain in more than one cluster. Using suffix tree algorithm, 3 clusters are retrieved as the time of posing the query on Google as shown in Fig 2).

The retrieved document snippets relevance weight $w(\ tf\ )$ for each labels can be obtained, calculating the weights, each $document\ d$ (snippet) in cluster labels, a weighted vector is constructed in relation to the super-concept, sub-concept node and relations as follows:

$$\vec{d} = (w_1, w_2, \dots w_n) \qquad (1)$$

$w_i\ is\ the\ weight\ of\ word\ (concept) i\ in\ document\ d$

$$w_i = \boldsymbol{tf} \qquad (2)$$

$where, Term\ Frequency\ (tf) is\ the\ times\ \ the\ word\ appearin\ the\ document$

The Domain Relevance scores (S) for each cluster labels would be found.Assume $C^k$ be the feature for concept K and $n_j$ be the collection of cluster labels that would be assign to concepts K, for each feature i$^{th}$ element of the concept K:

$$S = \frac{1}{n_j}\sum_{n_j\in k} W_{ij} \qquad (3)$$

Considering the leaf node concepts of Fig 1from each clusters, the feature $C^k$ for concept K are calculated based on super-concept, sub-concept and relation of concepts in the ontology hierarchy pattern extraction (X is a relation of Y or X and Y are sub-concepts of Z)

Assume K= 4

$C = \{C_1, C_2, C_3, C_4)\{superconcept[e], concept[e], subconcept[e], conceptrelation[e]\},$

$$c_i \in \{cluster[e]\}\ and\ \ c_k \in [cluster[c_j]]\}$$

$$S_i^k = \sum_{j=1}^{4}\frac{1}{n_j}\sum_{k=1}^{n_j} c_j\, Rel(c_i, c_k)\, W_{ij} \qquad (4)$$

Where $C$ is the correlation coefficient if the concept is in the ontology =1 else is 0 and

*Rel* is the relation between concept and its other concepts.

$n_j$is number of clusters of each leaf node entity defined in $C$

Moreover, some of the cluster labels candidates tends to overlap but the abstract entity that best matches the word or phrases is selected and other word or phrase that could not relate to the parents (super-class) or class concept in the hierarchy are ignored Therefore, from the clusters shown in fig2, using eqn (4) to calculate the feature of each cluster, it shows that the cluster 3 best described the query in relation to the ontology concept hierarchies (Bivalves). The score depends on features of the concept, each feature contributes independently to the score, unknown and unrelated features make no contribution to the score as in cluster 1. However, the cluster with the highest score is chosen as the cluster that best matches the concept (query). The system transform a feature represented document into concept represented one with the support of hierarchical structures of the ontology.

**Cluster1: Clams (27 docs, score: 16)**

[ 1] CLAMS | Cape Libraries Automated Materials Sharing
http://info.clamsnet.org/
Online catalog and resource sharing of the area libraries. Offers account access and links to area resources.

[ 2] CLAMS Web Catalog

[ 3] Giant Clam - National Geographic
http://animals.nationalgeographic.com/animals/invertebrates/giant-clam/
Learn all you wanted to know about giant clams with pictures, videos, photos, facts, and news from National Geographic.

[ 4] Clams | Washington Department of Fish & Wildlife
http://wdfw.wa.gov/fishing/shellfish/clams/
Clams belong to the group of animals called bivalves, which includes clams, mussels, oysters, and scallops. The soft body parts of these animals are enclosed

.
.
.

[ 9] Clam- Enchanted Learning Software
http://www.enchantedlearning.com/subjects/invertebrates/bivalve/Clamprintout.shtml
Clam Printout. Clams are invertebrates that burrow under the sea floor. These bivalves have two hard shells and a soft body.

.
.
.

**Cluster2:Clams Casino (6 docs, score: 8.29)**

[ 5] clammyclams's sounds on SoundCloud - Hear the world's sounds
http://soundcloud.com/clammyclams
clammyclams Clams Casino, NJ, United States. Follow Share a track Share. Facebook Twitter Tumblr Email WordPress StumbleUpon Blogger MySpace. Less.

[ 7] Clams Casino Official | Facebook
https://www.facebook.com/clammyclams
Clams Casino Official. 98562 likes · 176 talking about this. "What's remarkable about Volpe is that he basically helped invent the most interesting new ...

.
.
.

[31] Clams Casino – Free listening, videos, concerts, stats and pictures ...
http://www.last.fm/music/Clams+Casino
Watch videos & listen free to Clams Casino: I'm God, Natural & more, plus 49 pictures. Clams Casino is the pseudonym of New Jersey resident Mike Volpe, who ...

**Cluster3:Bivalves (4 docs, score: 2)**

[ 0] Clam - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Clam
The term clam generally refers to those bivalve molluscs that live buried in sand or silt, many of which are edible. Clams, like most molluscs, also have open ...

[ 4] Clams | Washington Department of Fish & Wildlife
http://wdfw.wa.gov/fishing/shellfish/clams/
Clams belong to the group of animals called *bivalves, which includes clams, mussels, oysters, and scallops*. The soft body parts of these animals are enclosed
...

[ 9] Clam- Enchanted Learning Software
http://www.enchantedlearning.com/subjects/invertebrates/bivalve/Clamprintout.shtml
Clam Printout. Clams are invertebrates that burrow under the sea floor. These bivalves have two hard shells and a soft body.

[20] clam - Wiktionary
http://en.wiktionary.org/wiki/clam
A bivalve mollusk of many kinds, especially those that are edible; as, the long clam (Mya arenaria), the quahog or round clam (Venus mercenaria), the sea clam ...

**Fig2: A small fragment of the Clusters from Google with STC**

## 4.3 Feature Vector

A formal model is presented that the document is represented as a vector of sentences which are composed of a vector of [15]. From eqn. (1), feature vector weight can be found by modifying Sureka et al, (2012) by adding concepts weight of the leaf nodes, mapping each node into a unique dimension of a M-dimensional vector space, computing concept weights taking threshold ($\emptyset$) = 0.45

Each document $d$ is considered to be vector in the M-dimensional feature term space. In this paper, the eqn. 6 weighting scheme is employed in which each document $d$ can be represented as follows:

$$\vec{d} = \{w(1,d), w(2,d), \ldots . w(M,d)\}$$

$$\boldsymbol{w_i = tf * correlation\ coefficient}\ (\alpha) + \boldsymbol{Prob. of\ concept} \qquad (5)$$

$where,$

$tf$ is the number of each concept (query) appeared in the snippet

$correlation\ coefficient\ is\ 1\ \ if\ the\ concept\ is\ in\ the\ ontology$

$otherwise =\ 0$

$$P(concept) = \frac{num\ of\ occurence\ of\ the\ concept(n_i)}{number\ of\ occurences\ of\ all\ the\ concept\ (N)}$$

$N$ is the total number of document in the collection corpus =30

$n_i$ is the number of document assigned to ith term (concept cluster), = 4 $n_i$can be document assisgned to concept hierarchies $(n_j, c_s,\ c_r)$ .

$$C_i^k = \frac{\sum_{n_j \in K} w_{ij}}{n_j} + \frac{\sum_{c_s \in K} w_{is}}{c_s} + \frac{\sum_{c_{Sb} \in K} w_{ir}}{c_{sb}} + \frac{\sum_{c_r \in K} w_{ir}}{c_r} \qquad (6)$$

For instance, ‘clam’ appear 2 times and bivalves is also 2 and other concept as well in snippet 1 of the cluster 3

Using eqn (4) for web search queries, we have the weight of the following concepts

$$d_1 = \frac{2 + \frac{4}{30}}{4} + \frac{1 + \frac{4}{30}}{4} + .......$$

$$d_1 = 0.533 + 0.283 + .....$$

$$d_1 = 0.82$$

$d_1$ = (clams, 0.82), (Bivalve, 0.56), (mussels, 0.53), (Oysters, 0.53)
$d_2$ = (clams, 0.71),( mussels, 0.83),( Oysters, 0.61) (Bivalve,0.56)
$d_3$ = (clamss, 0.82), (mussels, 0.55), (Bivalves, 0.78), (Oysters, 030)
$d_4$ = (clam, 0.52), (mussels, 0), (Oysters, 0.30), (Bivalves, 0.71)

## 5.0 Evaluation and Conclusion

From the retrieved document, it shows that the direct search using bag of words with vector space model (the classic information retrieval method) retrieved more irrelevant document. Semantic information retrieval method exploited the advantages of the semantic web to retrieve the relevant document related to the query. In the semantic search it considers the meaning for that submitted query and it displays the related results for that query. The performance of the STC algorithm was analysed using precision and recall. Precision (P) is defined as the proportion of retrieved documents that are relevant and where Ra is the relevant document retrieved and A is the retrieved document.

$$P = \frac{Ra}{A}$$

Recall (RR) is defined as the proportion of relevant documents that are retrieved and where R is relevant document.

$$RR = \frac{Ra}{R}$$

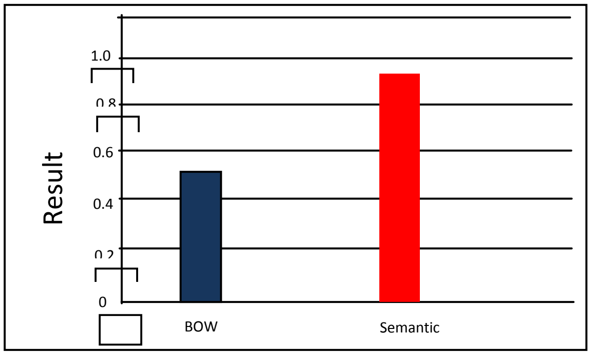

**Fig3: Shows Comparison between the BOW Search and the Semantic Search**

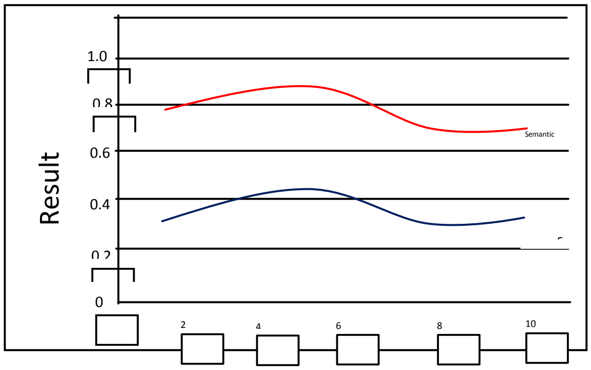

**Fig 4: Accuracy chart forthe BOW Search and the Semantic Search**

In conclusion, the proposed framework uses snippets to strongly calculate the feature vectors of web retrieved documents. The feature vectors are orthogonal, the dimensionality of feature vector is usually large, ontology can be used to reduced the dimensionality, therefore different similarity measures can be applied as the future work. The approach improves result that lead to retrieving more relevant documents with little irrelevant documents, therefore users are able to have a faster access to their required information. The improvements on the *ontology-based* indicate that the weighted suffix tree document clustering model tends to be a highly accurate documents clustering approach.

## 6.0 References

[1] 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 Online at: http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html.

[2] Salton G. and Buckley C. (1988): "Term-weighting approaches in automatic text retrieval," Inf. Process. Manage., vol. 24, no. 5, pp. 513–523.

[3] Carpineto, C., Osinski, S., Romano, G.,Weiss, D (2009): A survey of web clustering engines. ACM Computing Surveys 41(3), pp. 1–38.

[4] Hotho A., Staab S., and Stumme G. (2003): "Wordnet improves text document clustering". In Proceedings of the SIGIR Semantic Web Workshop, Toronto, Canada.

[5] Aitken S. and Reid S. (2000): "Evaluation of an Ontology-Based Information Retrieval Tool", Proceedings in ECAI, Berlin, Germany.

[6] Pan, X.S. (2002).: "A context-based free text interprete,". California Polytechnic State University San Luis Obispo Master's Thesis - Computer Science Department. Aug.

[7] Hanh H H. and Tjoa. A. M. ( 2006): "The state of the art of ontology-based query systems": A comparison of existing approaches. In Proceedings of ICOCI06 - The IEEE International Conference.

[8] Jain K., Murty M. N., and. Flynn P. J.( 1999): "Data clustering: a review". ACM Computing Surveys (CSUR), 31(3):264–323.

[9] Kruse P.M.,, Naujoks A., Roesner D., and Kunze M.( 2005): "Clever Search: A WordNet Based Wrapper for Internet Search Engines," In Proc. of 2nd GermaNet Workshop 2005, arXiv:cs/0501086v1.

[10] Tomassen, S. L. (2008): Searching with document space adapted ontologies. In Proceedings *1st world* summit on The Knowledge Society: Emerging Technologies and InformationSystems for the Knowledge Society, Athens, Greece, pp. 513-522.

[11] Tar, H.H. and Nyaunt, T.T.S. (2011): Ontology- Based Concept Weighting for Text Documents, World Academy of Science, Engineering and Technology Vol:5, 237-241.

[12] Sureka V and.Punitha S.C, (2012): "Approaches to Ontology Based Algorithms for Clustering Text Documents"Int.J.Computer Technology & Applications,Vol 3 (5), 1813-1817.

[13] Zamir O., and Etzioni O. (1998): "Web Document Clustering: A Feasibility Demonstration," Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 46-54.

[14] Shahnaz, F., Berry , M.W., Pauca, V.P. and Plemmons, R.J. (2006): “Document clustering using nonnegative matrix factorization”, Information Processing and Management, Vol. 42, Pp. 373–386.

[15] Kamel M. S and Hammouda K. M.,. (2004) “Efficient Phrase-Based Document Indexing for Web Document Clustering,”IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 10, pp. 1279-1296.