



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijircce.com](http://www.ijircce.com)

Vol. 5, Issue 3, March 2017

## Web Search Engine Using Ontology Learning

Seema Redekar, Vishal Chekkala, Siddhapa Gouda, Swapnil Yalgude

Assistant Professor, Dept. of I.T., SIES Graduate School of Technology, Nerul, Navi Mumbai, India

B.E Student, Dept. of I.T., SIES Graduate School of Technology, Nerul, Navi Mumbai, India

B.E Student, Dept. of I.T., SIES Graduate School of Technology, Nerul, Navi Mumbai, India

B.E Student, Dept. of I.T., SIES Graduate School of Technology, Nerul, Navi Mumbai, India

**ABSTRACT:** As the amount of information on the web is increasing at a faster rate, it is difficult to develop a search engine that provides efficient search by retrieving high quality documents related to query. Traditional search engines do not analyze the content of webpage and does not understand the meaning of user query. Thus, search engines should be enriched with methodologies that analyzes the content of webpages and provide more relevant results corresponding to users query. The proposed system uses ontology learning to enhance efficiency of search engine and uses only Wikipedia data, as it the largest repository that contain data from multiple domains. Ontology learning helps to determine the semantic relations. Semantic search helps the search engine to understand the user's queries. First, articles related to users query are retrieved and contents of the articles are analyzed using various algorithms to re-rank the webpages based on semantic similarity relation that exists among them. The proposed model provides a better approach to re-rank webpages than the traditional search engine.

**KEYWORDS:** Ontology learning, search engine, semantic search, Web page ranking

### I. INTRODUCTION

The amount of webpages and documents on the World Wide Web is increasing rapidly, it is almost impossible to get relevant documents without using a search engine. A search engine is an application for searching through documents on the web where keywords are given by the user [1]. Limitations of current search engines are they are keyword based search engines and do not understand what the users query mean. They see users query as string of characters and based on keyword matching they retrieve the webpages without understanding what they mean. Semantic Web can improve this situation by making the search engines understand what the content of the web pages are. If machines could understand what users query mean, they can do intelligent searching of the data corresponding to users query. If web pages are described with data that clearly identify the main concepts, then semantic search engines can find documents precisely related to users query. The proposed model uses ontology learning to enhance the efficiency of search engine by semantically understanding the relations between the concepts. In this paper, we provide a solution for machines to process data semantically. We use ontology learning methodologies to semantically model the significant concepts of a query along with its weighted semantic relations to other related concepts. The resulting ontology can be viewed as a signature of a topic that can be used to classify or re-rank document(s) based on the degree of similarity to the original query signature. [3]

### II. RELATED WORK

Approaches of current web search engines can be classified in to two groups: ontology based and Non-ontology based. Non-ontological based approaches are Hypertext Induced Topic search (HITS) and Page rank. HITS is a link analysis algorithm to rate Web pages. It is an iterative algorithm developed to quantify web pages value as hub and authority. The premise of the algorithm is that a web page serves two purposes: to provide information on a topic, and to provide links to other pages giving information on a topic. So it categorizes a web page in two ways:



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijircce.com](http://www.ijircce.com)

Vol. 5, Issue 3, March 2017

**Authority:** pages that provide important and trustworthy information on a given topic. So an authority is a page that is pointed by many hubs.

**Hub:** pages that contain links to authorities" i.e pointing to many pages. In HITS algorithm, ranking of the web page is decided by analyzing their textual contents against a given query. After collection of the web pages, the HITS algorithm concentrates on the structure of the web only, neglecting their textual contents.

PageRank algorithm uses the link structure of the web to determine the importance of a webpage. In this, a page obtains higher rank if sum of its back-links is high. This algorithm is based on random surfer model. The random surfer model assumes that a user randomly keeps on clicking the links on a page and if she/he get bored of a page then switches to another page randomly. Thus, a user under this model shows no bias towards any page or link. PageRank(PR) is the probability of a page being visited by such user under this model. For each web page, Page Rank value is pre-computed. For this over 25 billion web pages on the WWW are considered to assign a rank value [14].

The main drawback of such approaches is that user still required to browse through long list of Web pages to select those that are actually considered to be of interest. However, users usually check only one or two pages of the results returned by the search engine.

In recent years, development of ontology has become very common in World Wide Web in different industries such as medicine, electrical, oil and gas, etc. Most of the usage of Ontologies in industries is to categorize and organize the products to be ready to be use. In fact, the motivating factors to develop Ontologies are listed as follows:

To share common understanding of structure of information among people or software agents, to enable reuse of domain knowledge, to make domain assumptions explicitly, to separate domain knowledge from operational knowledge, to analyse the domain knowledge. Ontology is required to develop a semantic search engine [1].

Ontology-based search engines like Swoogle, OntoSelect, and OntoKhoj indexes ontologies that capture concepts, their properties, and their relationships for specific domain which can be used by computers to process within the data of those domains semantically. The foregoing ontology-based search engines are viewed as ontology libraries, they index an increasing number of ontologies, providing ranked ontologies to users based on his query. However, Current ontology ranking approaches mainly adopted methods used in conventional Web search engines to rank ontologies based on either popularity or based on some statistical measurements. Popularity-based approaches of ranking will not work for a large number of existing ontologies because of their poor connectivity and lack of referrals from other ontologies. Such 'self-contained' or 'isolated' ontologies would certainly receive poor rank, thus highlighting the need for additional ranking methods. Extracting ontology from the Web is a challenging task. One way is to engineer the ontology by hand, but this is expensive, tedious, and error-prone. Another problem with ontology-based approaches which is one of the biggest challenges that reside at the heart of numerous information processing applications is the Ontology matching problem which is the problem of finding the semantic mappings between two given ontologies [3].  
Web Ontology Language-OWL: Both OWL and RDFS are W3C recommended standards for modeling of Ontologies. OWL is on top of RDFS, which provides limited expressive means and it is not designed to present the complex knowledge. The main point to design OWL was not only to find a reasonable balance between expressivity of language, but also efficient reasoning i.e. scalability. In order to give the user a choice between different degrees of expressivity, three sub-languages of OWL-species of OWL have been designed expressed as: OWL Full, OWL DL, OWL Lite. OWL Full contains OWL DL and OWL Lite. OWL DL contains OWL Lite. OWL Lite is less expressive than OWL Full and OWL DL, but OWL Full is very expressive [1].

### III. INTRODUCTION TO ONTOLOGY LEARNING

Ontology learning (ontology extraction, ontology generation, or ontology acquisition) is the automatic or semi-automatic creation of ontologies, including extracting the corresponding domain's terms and the relationships between those concepts from a corpus of natural language text, and encoding them with an ontology language for easy retrieval. Ontologies are used to capture knowledge about domain of interest and also the relationships that hold between them

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijircce.com](http://www.ijircce.com)

Vol. 5, Issue 3, March 2017

[11] .An Ontology should be used to solve the semantic issues and share knowledge with and among computers. An Ontology supports sharing information and knowledge, defining the relationships between different resources, understanding of the domain. [12].Ontologies have a main role in Semantic Web Vision. The main goal of Ontologies is to give conceptual description of a domain and models the domain with their concepts, relations and properties. Defining the complex reasoning will be possible by the ontology concept[1].Ontology Learning from Text mostly focuses on the automatic or semi-automatic generation of lightweight taxonomies by means of text mining and information extraction. Many of the methods used in ontology learning from text are inspired by previous work in the field of computational linguistics, essentially designed in order to facilitate the acquisition of lexical information from corpora. Such ontology population methods derive facts from text. [13]

## IV. PROPOSED SYSTEM

The proposed system presents a method for extracting ontology concepts from text and webpages. In the proposed system, a semantic graph is built for any topic which indicates the relatedness between extracted concepts. Finally, it generates a list of concepts which are ranked based on their relatedness to the initial search topic. The proposed system comprises of stages namely Extract Keyphrases, search, crawl and download Wikipedia articles, computing similarity of articles (semantic similarity), Clustering the semantic graph, Computing similarity between clusters, Ranking the clusters. The system architecture is shown in figure 4.1

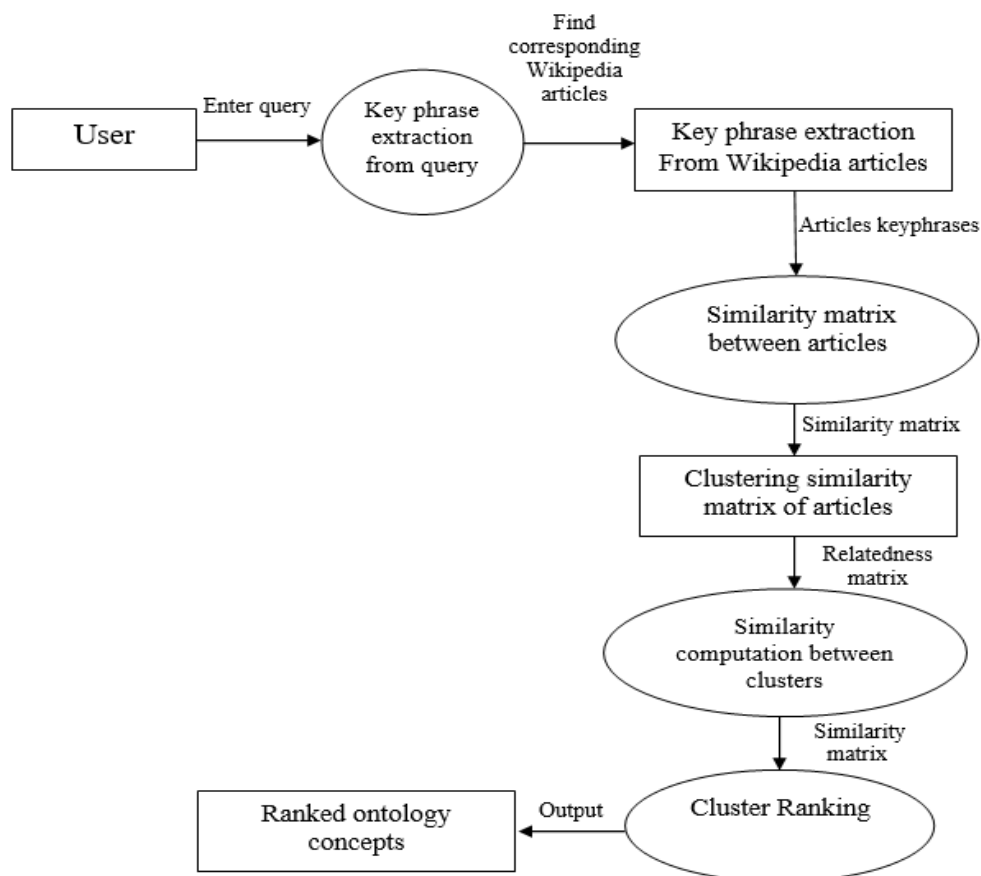


Figure 4.1: System architecture



# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijircce.com](http://www.ijircce.com)

Vol. 5, Issue 3, March 2017

## 1] Extract Keyphrases:

Keyphrases are group of phrases that represent a document (eg: Electronics engineering). Keyphrases are extracted so that they provide meaning to the users query. As user understands an article based on keyphrases present in it, the proposed model extracts keyphrases from users query to make the search engine understand what the users query represents. Unsupervised keyphrase extraction algorithm is used in proposed model. [2][4][3].

## 2] Search, Crawl and download Wikipedia articles:

Due to large size of offline Wikipedia (around 5TB) and not able to retrieve the most updated data, it is better to use Wikipedia API and retrieve data online. Based on user query, Wikipedia articles are crawled and downloaded respectively. Keyphrase extraction is again used and applied on the articles and is used for further processing. [3]

## 3] Computing similarity of articles:

Semantic Similarity between the articles can be computed by number of common keyphrases the articles share with each other. In proposed model, every article is compared with every other article that are retrieved. In proposed model, Cosine similarity is used to compute the similarity between articles using the following equation.

$$\text{Similarity (A, B)} = \text{Cosine } (\Theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

Where A and B are vectors that represent keyphrases of articles A and B. The result is cosine angle between the two vectors. The values of result will be between 0 and 1 where 0 represents that articles are not related to each other at all and 1 represents articles that are highly related to each other. The output of this step is  $N \times N$  matrix (semantic graph), where N is number of Wikipedia articles. [3] [5].

## 4] Clustering the semantic Graph:

As the number of communities will be different for different similarity matrix (semantic graph), so clustering techniques where pre-defined number of clusters are required cannot be used in our model. In the proposed model, a clustering algorithm (Table1) [3] is developed based on Ant Colony Optimization algorithm (ACO) [6] to detect different communities in the semantic graph. Ant colony optimization works by taking those nodes in the cluster which have score greater than or equal to the average score i.e fitness. For example, assume ant starts by first visiting node A with similarity score 0.4, then node B with similarity score 0.6, then the fitness can be calculated by averaging the similarity scores of visited nodes  $(0.4 + 0.6) / 2 = 0.50$ . The resulted value will be used as a threshold for next candidate node to be visited which should have similarity score greater than or equals to the average score. This property is very useful, as nodes that are highly related to each other will be grouped together; and unrelated nodes will be removed [3].

## 5] Computing similarity between clusters:

After applying Ant colony optimization and generating clusters, we now compute the similarity between these clusters. As clusters that are related to each other will share common topics. Similarity computation between clusters is computed by applying Dice-coefficient using the following equation:

$$D = \frac{2|X \cap Y|}{|X| + |Y|}$$

Where X, and Y represents two clusters. The output is the similarity D between two clusters X, and Y with value range from 0 (completely unrelated) to 1 (identical). Each cluster is paired for similarity computation with other clusters using the previous equation. The output of this step is  $N \times N$  matrix (graph) of similarity between clusters where N is the number of clusters. [3]

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijircce.com](http://www.ijircce.com)

Vol. 5, Issue 3, March 2017

**Table 1. Ant Cluster Algorithm [3]**

**Inputs**

Similarity Matrix:  $N \times N$  Similarity Matrix of Wikipedia Articles

**Output**

Relatedness Matrix:  $N \times N$  Relatedness matrix of Wikipedia Articles

**Initialize**

1. For each ant:

    Ant fitness 0; // average value of nodes visited  
    Nodes visit Null; // List of nodes that ant visits

**Operation**

2. For every row in Similarity Matrix

    2.1 Select starting nodes where ants will start their tours from

3. For each ant

    3.1 Update Nodes visit current ant position[x,y];

    3.2 Update Ant fitness value of current position[x,y]

**4. Ant Tour**

    4.1 Select next node in ant tour based on :

        4.1.1 select the neighbour node with the highest similarity score to the current node IF its similarity score greater than or equal to the ant fitness value

**If** next node exist in its tour list; **End** ant tour;

**Otherwise**;

**Update** the relatedness score of visited node with the value in Current ant position[x,y] in similarity matrix

    4.2 Update ant tour list

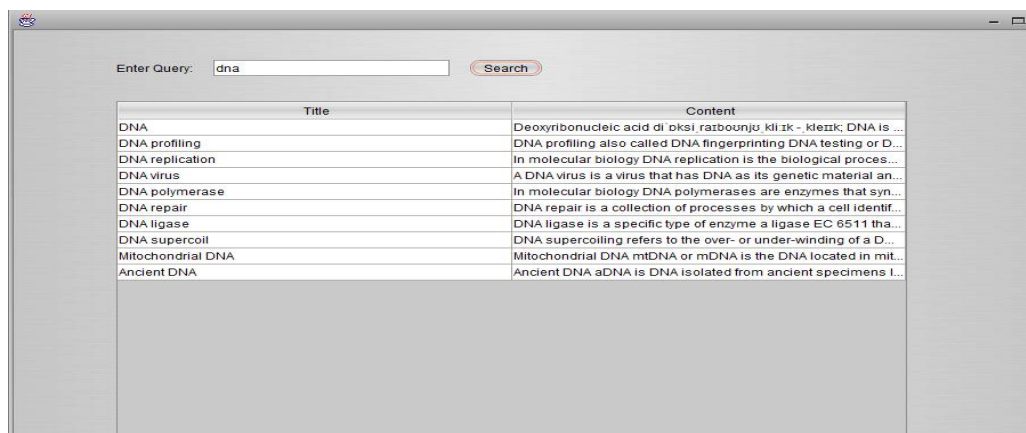
    4.3 Calculate and Update Ant Fitness

    4.4 Go to step 4.1

6) Ranking clusters:

This step involves extracting and ranking the related clusters and remove the unrelated clusters. Ant cluster algorithm is used again given the similarity matrix between the clusters as input. Rank of clusters is emerged during execution of ant cluster algorithm. The clusters with highest average scores will be on top of the results i.e in descending order of average scores, clusters will be ranked. For example, Wikipedia title 1 is cluster 1, Wikipedia title 2 is cluster 2, Wikipedia title 3 is cluster 3 and so on, assume that if cluster 4(Wikipedia title 4) has highest average score than the rest of the clusters, then it will be ranked 1 in the output of the results. The output of this step is ranked concepts i.e cluster name which is title of Wikipedia article.

## V. SIMULATION RESULTS



Title	Content
DNA	Deoxyribonucleic acid di oksî razbounjû klîzk - klezk; DNA is ...
DNA profiling	DNA profiling also called DNA fingerprinting DNA testing or D...
DNA replication	In molecular biology DNA replication is the biological proces...
DNA virus	A DNA virus is a virus that has DNA as its genetic material an...
DNA polymerase	In molecular biology DNA polymerases are enzymes that syn...
DNA repair	DNA repair is a collection of processes by which a cell identif...
DNA ligase	DNA ligase is a specific type of enzyme a ligase EC 6511 tha...
DNA supercoil	DNA supercoiling refers to the over- or under-winding of a D...
Mitochondrial DNA	Mitochondrial DNA mtDNA or mDNA is the DNA located in mit...
Ancient DNA	Ancient DNA aDNA is DNA isolated from ancient specimens I...

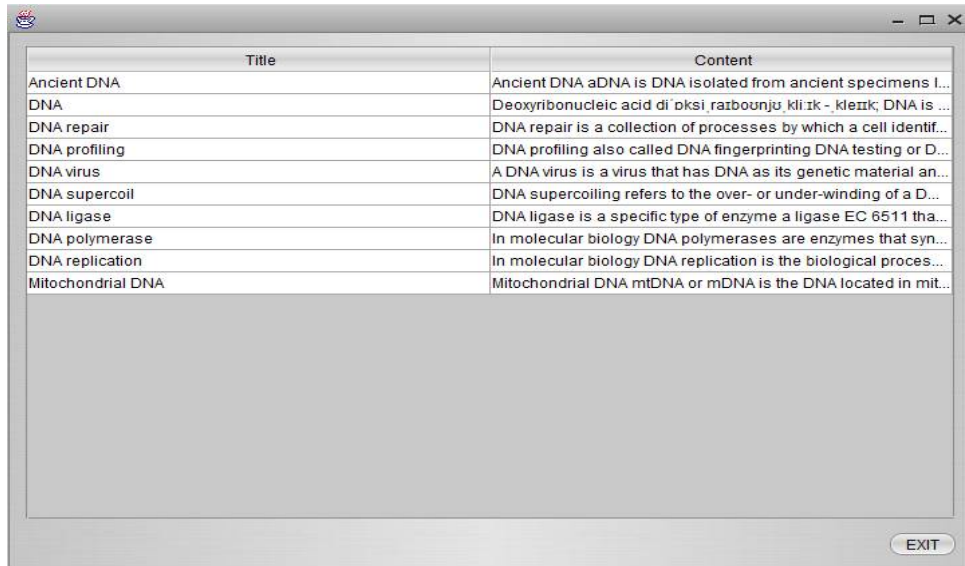
**Figure 5.1: Results of query “dna” that are retrieved from Wikipedia**

# International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: [www.ijircce.com](http://www.ijircce.com)

Vol. 5, Issue 3, March 2017



Title	Content
Ancient DNA	Ancient DNA aDNA is DNA isolated from ancient specimens I...
DNA	Deoxyribonucleic acid di'oksi razbounjʊ kli:ɪk - kleɪk; DNA is ...
DNA repair	DNA repair is a collection of processes by which a cell identif...
DNA profiling	DNA profiling also called DNA fingerprinting DNA testing or D...
DNA virus	A DNA virus is a virus that has DNA as its genetic material an...
DNA supercoil	DNA supercoiling refers to the over- or under-winding of a D...
DNA ligase	DNA ligase is a specific type of enzyme a ligase EC 6511 tha...
DNA polymerase	In molecular biology DNA polymerases are enzymes that syn...
DNA replication	In molecular biology DNA replication is the biological proces...
Mitochondrial DNA	Mitochondrial DNA mtDNA or mDNA is the DNA located in mit...

Figure 5.2: Results of Ranked Ontology Concepts for query “dna”

## VI. CONCLUSION

Conventional search engines are text based, i.e. matches only the keywords and does not understand the content of web pages and returns the result. The proposed system analyzes the content of web page and uses ontology learning to enhance the efficiency of search engine by making the search engine understand the content of web pages. This is done by semantically extracting the relationships between the concepts that represent the query. It allows highly relevant pages corresponding to users query to be on top of results and also solves the information overload problem.

## REFERENCES

- [1] NooshinAghajani, “Semoogle - An Ontology Based Search Engine”, In June 2012.
- [2] Aliaa A.A. Youssif, Atef Z. Ghalwash, and Eslam Amer, “KPE: An automatic keyphrase extraction algorithm”, IEEE proceeding of International conference on Information Systems and Computational Intelligence (ICISCI 2011), pp.103-107, 2011
- [3] Eslam Amer, “Enhancing efficiency of Web search engines through ontology learning from unstructured information sources”, In 2015 IEEE 16<sup>th</sup> International Conference on Information reuse and Integration.
- [4] Sifatullah Siddiqi , AditiSharan, “Keyword and Keyphrase extraction techniques: A literature review”, In 2015 IJCA Volume 109-No.2
- [5] Anna Huang, “Similarity measures for Text Document clustering”, University of Waikato, Hamilton, New Zealand.
- [6] Leandro de Castro, “Immune, Swarm, and evolutionary algorithms Part 2: Philosophical Comparison”, International conference on Neural Information Processing), Workshop on Artificial Immune systems, vol. 3, pp. 1469-1473, 2002.
- [7] A. Maedche, S. Staab, “Ontology learning for the semantic Web”, IEEE Intelligent Syst. 16 (2) pp.72-79, 2001.
- [8] Aliaa A.A. Youssif, Atef Z. Ghalwash, and Eslam A. Amer. “HSWS: Enhancing efficiency of web search engine via semantic web”. ACM Proceeding of the 3rd International Conference on (MEDES'11), pp. 212-219, 2011.
- [9] Maryam Hazman, Samhaa R. El-Beltagy, and Ahmed Rafea. “Ontology Learning from Textual Web Documents”. Proceeding of INFOS2008, pp. 113-120, 2008.
- [10] J. Hein and J. Hendler. “A Portrait of the Semantic Web in Action.” IEEE Intelligent Systems, 16(2), pp. 54-59, 2001.
- [11] M.Horridge, H. knublauch, A. Recto, R.Stevens, and C.Wroe, “A Practical guide to building owl ontologies using the protégé-owl plugin and co-ode tools,” vol 27, pp. 0-117, 2004.
- [12] M.Szlezak, “Markup language in realization of design tasks,” 2005 [Online].
- [13] Jens LEHMANN and Johanna VOLKER, “An Introduction to Ontology learning”.
- [14] Pooja Devi, Ashlesha Gupta, Ashutosh Dixit., “Comparative study of HITS and PageRank link based Ranking Algorithms”, International Journal of Advanced research in Computer and communication Engineering, Vol. 3, Issue 2, February 2014