ONTOLOGY BUILDING FOR SEMANTIC SEARCH ENGINE

Anjana Anilkumar[1], Itiksha P. Lade [2], Maithili V. Bhide[3], Ancy Gonsalves[4]
[1](Department of Computer Engineering, University of Mumbai, Palghar
Email: anjana07k@gmail.com)
[2](Department of Computer Engineering, University of Mumbai, Palghar
Email: lade123itiksha@gmail.com)
[3](Department of Computer Engineering, University of Mumbai, Palghar
Email: maithilibhide1996@gmail.com)
[4](Department of Computer Engineering, University of Mumbai, Palghar
Email:gonsalvesancy@gmail.com)

ABSTRACT

The important aspect of web is search engine. To search relevant information the search engines are used. The problems faced by normal search engines gave rise to semantic search engine. Search engine acts as a tool to find the data. Our work aims at building a semantic search engine based on ontology. Semantic search engine interprets a query in a structured format and provides the users with meaningful data which have relations among the terms which are easily understandable. Ontology is a keyword based approach to map behavior of an object to identify and analyze its traits. Our purpose is to increase the chances of getting more accurate results. Semantic web is a concept that returns results relevant to the searched query. Our system is mainly suitable for primary school students to get precise, efficient and understandable data about a particular domain.

Keywords*—Ontology, OWL, RDF, Semantic Search, SPARQL, Triplet Extraction,*

I. INTRODUCTION

Normal search engines give us links of the web pages that contain the query searched using the said search engine. Semantic search engine is a data searching technique in which a search query aims to provide the intend and contextual meaning of the words , what the user is looking for. The search engine needs the user to enter a query with which data, text, images and videos suitable and relevant to primary school students will be displayed as a output. Semantic web search is one of the most important aspects of customizing user experience to provide exact set of results in the case of ambiguous query processing. Modern search engines deploy semantic search to a great extent, but that is not to an extent where it can be used to help and optimize search. These restrictions are due to the fact that the search engines like google.com or yahoo.com are generalized and specialized search engines for different environmental needs like a code search engine for computer science student or a math search engine for math student. Here we propose to build a semantic search engine that could provide a specific set of optimized results for the domain of physics. As the user enters the search key, the first step is the lexical analysis process. In this stage a sequence of characters that is the search key is converted to a sequence of tokens. Now, the information retrieval is done using the process of stemming. Stemming is the process of reducing derived words to their root form.

Our system uses the concept of OWL. Web Ontology Language (OWL) is semantic web language designed to represent complex knowledge about things and relations between them. The consistency of knowledge is verified using OWL. The semantic searches are improved by RDF data. Resource Description Framework (RDF) stores web based data with well defined meaning. It is a standard model for interchanging data on web. RDF extends the linking structure of web to use URLs to name the relationship between things as well as the two ends of link. It is directed, labeled graph data format for representing information in the web. This specification defines the syntax and semantics of SPARQL query language for RDF. SPARQL is needed for accessing the data on ontology. The results of SPARQL queries are results sets and RDF graphs. Triplet extraction algorithm is applied on the text data which is fetched from various sources. The phrases from the triplets are compared for the similarity using phrase2vec model algorithm. The final aim of our work is to make learning enjoyable for kids.

II. LITERATURE SURVEY

Paper [1] proposes designing of the ontology based engine and then mapping the ontologies together. The data here is extracted from the ontology. The crawled information is added to the ontology by creating an OWL individual for each of them only if they aren't present in the knowledgebase. Next part mentioned in the paper is about extending the traditional full test indexed based on Lucene6 indices. The construction of Lucene index is such that each entry represents a soccer event. Firstly, the ranking of fields containing extracted data is boosted to stress importance of them. Secondly, the fields are re-ranked according to their importance. Hence, the part here is about ranking the extracted data. The retrieved document is pre-processed using pre-processing module. The next step which is mentioned in the paper is the computation process which has the pre-processed module as the input and extends the url list for further processing. User inspects the results of the above mentioned crawling process. Adding RDF metadata to the local system and refining the evolving ontology based on the analysis of documents is contained in the document list .The proposed paper concentrates on the semantic similarity and the whole process including query submission and information annotation. The data retrieval can give more precise answers to users without ranking document.

Paper [2] proposes a semantic based multiple web search engine in which each page possess semantic metadata that record additional details concerning the web page itself. The initial set of relations is firstly exploded from the query by adding hidden relations. The similarly ratio is then calculated. This method is to be applied on each property individually. The user specifies the relations. In this paper, lexical relations like synonyms, antonyms and homonyms between keywords have been used so that the query results can be expanded and the formal queries are automatically formulated. This is not scalable. This method focuses on capturing more metadata. An ontology dictionary is also been built.

Paper [3] proposes a conceptual architecture for the semantic search engine. They discuss the component required by engine and requirements of these components. They mainly focus on use of relational database to store the knowledge base. The data on the web is very huge therefore it is difficult for user to search the required information. The system uses ontology created in DAML or OWL (Web Ontology Language).PROLOG language is used for inference engine implementation. By using OWL, the language standardization problem is solved. In this, RDBMS is used which are multi-user and do not impose any limitations on the number of ontologies used. Two main inference engine used are CWM and EULER.CWM is implemented in python and uses RDF. Different components are used for both the servers. The tools used here are different for client and server side. The model overcomes the drawbacks of existing inference engine.

III. PROPOSED WORK

Ontology based semantic search engine is a search engine that gives relevant information about a particular domain. The domain which we consider is of physics. The system will give relevant information about only physics related searches. Ontology is a set of concepts and categories on a subject area or domain that shows their properties and the relations between them. Semantic search seeks to improve search accuracy by understanding the searcher's intent and the contextual meaning of terms as they appear in the searchable dataspace to generate more relevant results.

The user first enters the search key. The users for our system are the primary school going students. These sequence of characters which is entered in the searchable dataspace are converted to the sequence of token. This process of conversion is known as lexical analysis process. The derived words are reduced to their root form by the process of stemming. Different data sources when integrated arises a problem. To solve this problem triplet extraction process is used. Triplet extraction extracts triplet patterns from the user query. Then SPARQL query is generated. SPARQL is a language used for accessing the ontology. The data from the ontology is retrieved if the SPARQL query is generated. When the user query is submitted to the system, this query is needed to translate if the input query is in unstructured form, it will be difficult to define which word are subjects or objects. Hence, triplet extraction is used to identify the subjects, predicates and objects. Once the triplets are formed, the validity of phrase is checked. If no, then key based search is done and the result serves as the input key. If the phrase is valid then the next process is RDF mapping. Triplet extraction is important because if the input query is in unstructured form, it will be difficult to define which words are subjects or objects. Hence, Triplet extraction is used to identify the subjects, predicates and objects.

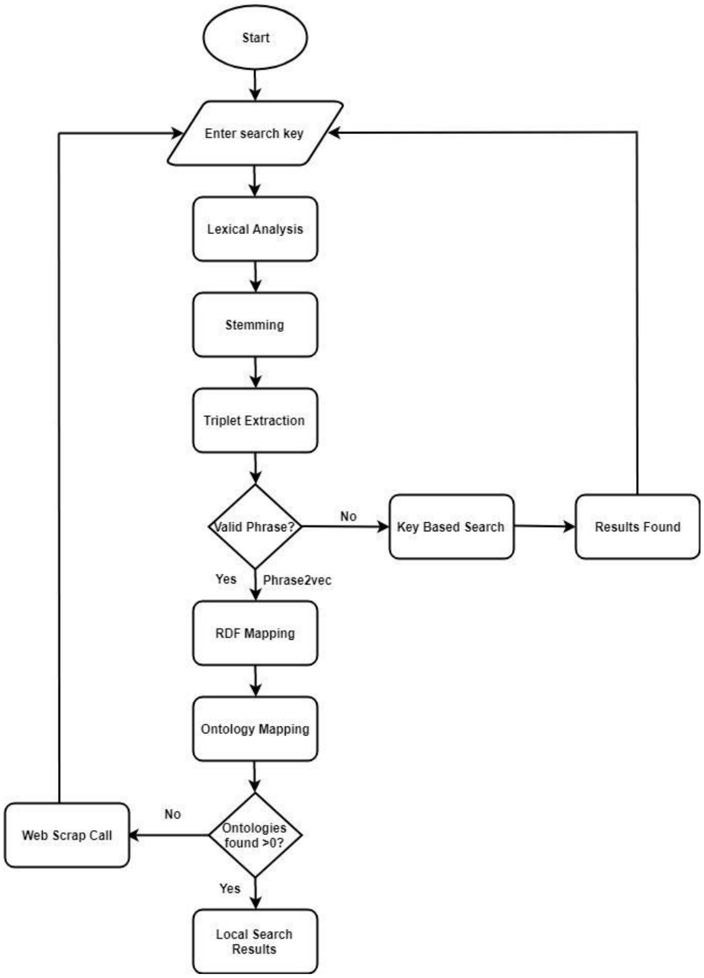


Fig 1. Flowchart of Semantic Search Engine

Once the triplets are formed, the validity of the phrase is checked. If no, then the key based search is done and the result serves as the input key. If the phrase is valid then RDF mapping is the next process. Resource Description Framework (RDF) is a family of World Wide Web Consortium specifications originally designed as a metadata data model. It basically describes the resources which have their own properties. It is any object which is uniquely identifiable by a Uniform Resource Identifier (URI).Each property is uniquely identified here. A query as an input can easily be understood by a human but it is difficult for machine to understand. Hence, RDF is to be used.

Once the meaning of the query if found, the information from different sources are interrelated. This is the ontology mapping also known as semantic integration. Our system then checks the rank of the ontologies that are found. If the data is found then the following information is displayed in the urls. If not, then web scraping module comes into picture. The system extends the search to different websites. If the relevant information is present in those links then the data is parsed. The proposed system uses OWL, Web Ontology Language (OWL).OWL is a family of knowledge representation languages for authoring ontologies.

A. Text Summarization :

Summarization is the process of extracting important information from the source text to present that information to the user. It is a task of Natural Language Processing (NLP).The text summarization that we apply in our work is extractive i.e, the summary extracts the important section from the original file reproducing the same words as that of the original file. It scans the most important sections or can say that it extracts the key points while calculating correlation,we need summary first. Ontology summarization is defined as the process of distilling knowledge from ontology to produce an abridged version for a particular user's need or task's requirement. When a user enters the search key, the search key is scanned from the summaries. It the particular entered search key is present in the summary then it's article is displayed to the user. The data can directly or indirectly refer to a phrase. For the phrases which are indirectly relatable, we need text summarization part.

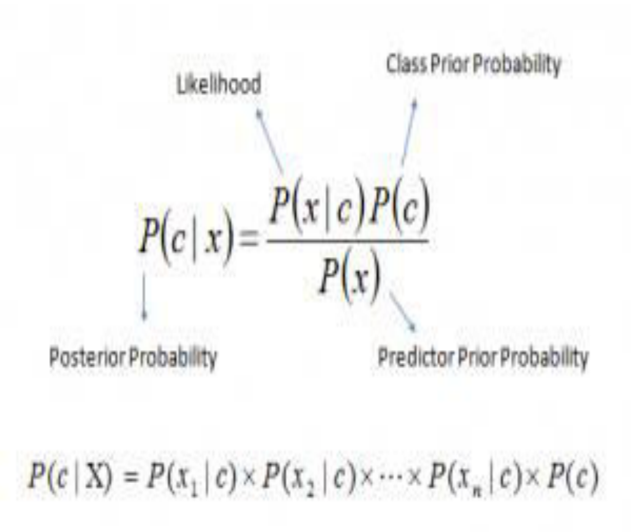
B. Naïve Bayes Algorithm :

Naive Bayes Algorithm is a classification technique based on Bayes theorem. It assumes that the presence of a particular feature is unrelated to the presence of any other feature. It is easy to build and particularly useful for very large data sets. This algorithm is also used for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.

Step 1: Convert the data set into a frequency table.

Step 2: Create likelihood table by finding the probabilities.

Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.



$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$$

$$P(c\mid X)=P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

C. *Triplet Extraction Algorithm :*

In this algorithm, the paragraph of text is separated sentence wise and noun-phrases which would later serve as the two entities bound by a relationship. This algorithm takes less execution time than the other algorithms and will be a useful algorithm for extracting triplets from the unstructured sentence.

The noun phrases which would later serve as two classes (entities) connected by a relationship and the verb phrase which would later serve as relationship is extracted by Triplet Extraction – Semantic Tree.

Triples mainly contain three components:

- **Subject** – The subject stands for the URL, document, the person we are talking about.
- **Predicate** – Relations are stored in Predicates.
- **Object** – Object holds that relation value for the particular subject.

A sentence (S) is represented by the parser as a tree having three children: a noun phrase (NP), a verbal phrase (VP) and the full stop (.). The root of the tree will be S. Firstly we intend to find the subject of the sentence. In order to find it, we are going to search in the NP subtree. The subject will be found by performing breadth first search and selecting the first descendent of NP that is a noun. Secondly, for determining the predicate of the sentence, a search will be performed in the VP subtree. The deepest verb descendent of the verb phrase will give the second element of the triplet. Thirdly, we look for objects. These can be found in three different subtrees, all siblings of the VP subtree containing the predicate. The subtrees are: PP (prepositional phrase), NP and ADJP (adjective phrase). In NP and PP we search for the first noun, while in ADJP we find the first adjective.

D. *Phrase2Vector Algorithm :*

In this algorithm, we begin by constructing a symmetric matrix of all the relations obtained from Triplet Extraction. We use the Phrase2Vec Model to find out the similarity between them. The measure of similarity is cosine similarity. We use the python module, genism which is designed to automatically extract semantic topics from documents, as efficiently (computer-wise) and painlessly (human-wise) as possible. The vectors within the set that need to be ignored are taken into consideration. In our case these words are the stop words. The centroid of these vectors is calculated and the phrase vector is obtained. A centroid is the appropriate measure because it incorporates equal contribution of each word that went into forming the semantic meaning of the phrase. So, if the similarity between two vectors is closer to one, they are more similar.

IV. RESULTS

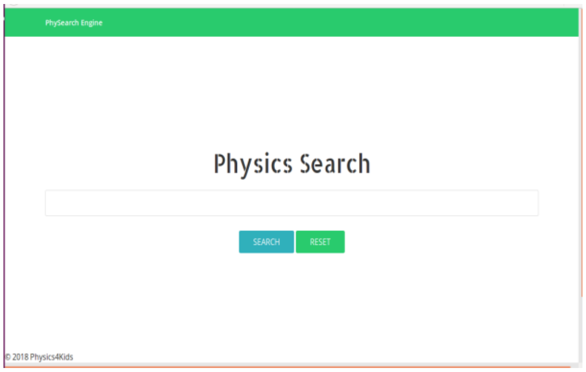


Fig 2. Home Page

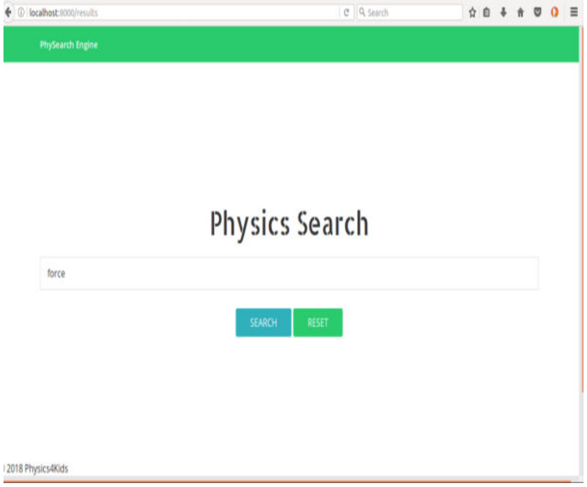
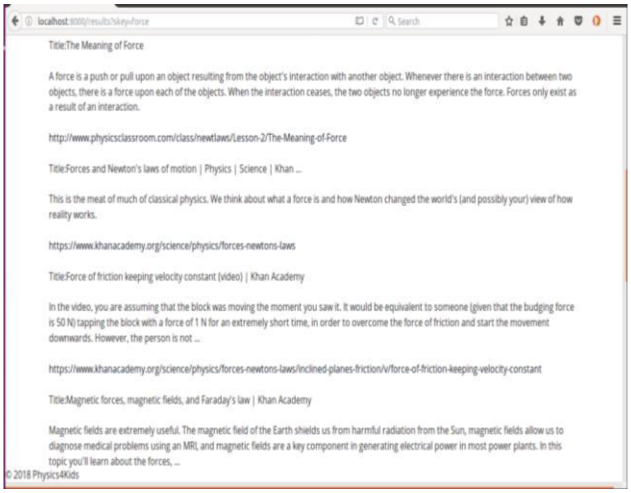


Fig 3. Search Query



Title:The Meaning of Force

A force is a push or pull upon an object resulting from the object's interaction with another object. Whenever there is an interaction between two objects, there is a force upon each of the objects. When the interaction ceases, the two objects no longer experience the force. Forces only exist as a result of an interaction.

http://www.physicsclassroom.com/class/newtlaws/Lesson-2/The-Meaning-of-Force

Title:Forces and Newton's laws of motion | Physics | Science | Khan ...

This is the meat of much of classical physics. We think about what a force is and how Newton changed the world's (and possibly your) view of how reality works.

https://www.khanacademy.org/science/physics/forces-newtons-laws

Title:Force of friction keeping velocity constant (video) | Khan Academy

In the video, you are assuming that the block was moving the moment you saw it. It would be equivalent to someone (given that the budging force is 50 N) tapping the block with a force of 1 N for an extremely short time, in order to overcome the force of friction and start the movement downwards. However, the person is not ...

https://www.khanacademy.org/science/physics/forces-newtons-laws/inclined-planes-friction/v/force-of-friction-keeping-velocity-constant

Title:Magnetic forces, magnetic fields, and Faraday's law | Khan Academy

Magnetic fields are extremely useful. The magnetic field of the Earth shields us from harmful radiation from the Sun, magnetic fields allow us to diagnose medical problems using an MRI, and magnetic fields are a key component in generating electrical power in most power plants. In this topic you'll learn about the forces, ...

© 2018 Physics4Kids

Fig 4. Displayed Results

V. CONCLUSION AND FUTURE WORK

We have implemented building an ontology which would form the base of a semantic search engine. The most important part of building this ontology lies in triplet extraction and finding out similarity between the relations. We presented a system that will help the school going students to get relevant information about a specific domain. The domain we considered here is of physics. This ontology based semantic search engines make the use of the search engine very easy and enjoyable for students. The user interface can be further improved by adding animations and symbols that would attract kids and also change the fonts and colors of the text. The idea is to make UI more attractive.

REFERENCES

[1] N.Vanjulavali, Dr.A.Kovalan ,“On Ontology based Semantic Search Engine ,”in international Journal of Computer Science and Engineering Technology, vol.2 , August 2012.

[2] Ms.S.Latha Shanmugavadivu, Dr.M.Rajaram ,“On Semantic based Multiple Web Search Engine ,”in international Journal of Computer Science and Engineering Technology, vol.2 , 2010.

[3] Qazi Mudassar Ilyar, Yang Zong Kai and Muhammad Adeel Talib ,“On A Conceptual Architecture for Semantic Search Engine ,”in IEEE, 2004.

[4] Arooj Fatima, Cristima Luca and George Wilson , “On New Framework for Semantic Search Engine ”, in IEEE, 2014.

[5] Wei-Dong Fang, Ling Zhang, Yan-Xuan Wang and Shou-Bin Dong , “On Toward A Semantic Search Engine Based On Ontologies ”, in IEEE, 2005.

[6] Leyla Zhuhadar, Olfa Nasraoui, Robert Wyatt and Elizabeth Romero, “On Multi-language Ontology-Based Search Engine ”, in IEEE, 2010.

[7] Zin Thu Thu Myint and Kay Khaing Win, “On Triple Patterns Extraction for Accessing Data on Ontology ”, International journal of Future Computer and Cmmunication, Vol 3, February 2014.

[8] Mehrnoush Shamsfard, Azadeh Nematzadeh and Sarah Motiee“An Ontology Based System for RankingDocuments”, International Journal of Computer Science, vol .1, 2006.

[9] S. Lu, M. Dong and F. Fotouhi, "The Semantic Web: Opportunities and Challenges for Next-Generation Web Applications", International Journal of Information Research, 7(4), 2002

[10] L. Ding, P. Kolari, Z. Ding, and S. Avancha, “Using Ontologies in the Semantic Web: A Survey,” Ontologies, 2007.