A Novel Approach for Ontology Focused Inter-Domain Personalized Search based on Semantic Set Expansion

[1]Sairam Haribabu, [1]P Siva Sai Kumar, [1]Sairam Padhy, [1]Gerard Deepak, , [1]A Santhanavijayan, [2]Naresh Kumar D

[1]Department of Computer Science and Engineering
[2]Department of Mechanical Engineering
[1,2]National Institute of Technology,Tiruchirappalli
Tiruchirappalli, India
sairamharibabu@gmail.com

***Abstract* – The Web is overloaded with information and its exploration and retrieval is quite tedious is a cumbersome task. In the era of Semantic Web, there is a requisite of Semantic strategies for recommending webpages. In this paper, a strategic semantic paradigm for Ontology Driven Semantic search has been proposed. The proposed scheme incorporates a set expansion mechanism for interdomain exploratory semantic search. The proposed scheme for searching semantically computes concept similarity between concepts from most similar domains, and the set expansion is based on these concepts. The ontologies are visualized using a triple store and personalization has been imbibed into the proposed scheme. The initial step requires determination of the domains and scope of the ontologies, followed by identification of classes, defining instances of each class and the relationship among them. Multiple user profiles will be created to personalize the search results according to the search history of each user. Through this model, the problem of irrelevant search results is reduced, and in the process, reduces the probability of going through numerous results as in case of a normal search engine. Also, the search engine is made scalable to any dataset irrespective of its content. An overall F-measure of 96.64% has been achieved.**

Keywords—Domain Scalable, Inter-Domain, Ontologies, Personalized, , Semantic Search, Triple Store Database

I. INTRODUCTION

The data in the World Wide Web is growing at a fast pace in the data repositories, thanks to various factors like users, sensors, systems and applications. For example, numerous transactions take place over the internet daily and the social media platforms like Instagram, WhatsApp, Facebook and Google+. The data is produced at such a rapid pace and in such huge volumes and formats that it has practically led to a competition among the developers to keep looking for a technique to obtain relevant results for searches and ignoring the irrelevant ones. With the ever-increasing volume, variety and velocity of data in the data repositories online, one pertinent problem that has been plaguing the software developers around the world, is that of extracting correct, precise and accurate information with minimal lag time. These issues can be countered by developing a technique that can successfully extract useful data and rectify the problems that exist in generating a semantic search over the World Wide Web (www).
Therefore, there is a need to identify these issues and present a semantic search engine which addresses issues like inter-domain querying, personalized results and optimized complex query retrieval. Ontology is a concrete solution to the problems of using varying terminology to refer to one concept or using the same term to describe varying concepts. Technological advancements in recent times have led to the invention of multiple smart devices – TVs, Smart phones, Tablets, Watches etc. In addition to these multiple devices that have taken mankind by storm, which led to the inevitable birth of several input methods. There have been a lot of advancements from inputting a word/sentence through keyboard to providing the same input through voice

Search engines have brought about a pleasant revolution in terms of improving search accuracy and presenting us with accurate and precise answers. This is accomplished by understanding the intent of the user behind asking that query, which is the job of a semantic search engine. A semantic search engine doesn't just perform search operations by evaluating the query on its face value, that is, keyword search, but searches according to the contextual meaning of the query, by taking the users' search history into consideration. The semantic search eliminates any ambiguity that arises during a search query, as it gets the real meaning of the terms, by virtue of a powerful semantic network that was created with the available data. So, for example, if a query reads "Who is the tallest man on earth?", which is followed by the query "How tall is he?", a non-semantic search engine won't understand that the 'he' in the second query refers to 'the tallest man on the earth', instead what it will do is retrieve all the web pages which have one or more occurrences of the phrase 'How tall is he?' and return them to the user. Or as another example, if you search for "Unicorn" after performing 20 searches on bikes, then the semantic search engine would understand that your intended search query refers to the bike, 'Honda Unicorn', rather than suggesting web pages about the fictional animal, unicorn.

***Motivation**:* Traditional keyword-based search engines fail to comprehend the meaning or the users' intent behind a query. The need for this advancement was a result of poor correlation of similar terms lead to undesirable results arising out of a search engine. Therefore, a search engine was needed which understands the semantics of the query rather than only the keywords itself. Moreover, a semantics driven Ontology based search engine is a mandatory requirement for extracting useful and relevant information

from the Semantic Web 3.0, which is the evolving standard for the World Wide Web (WWW).

Contribution: The proposed system is a search engine capable of deducing the semantics of a query by incorporating ontologies. It has the capability to include ontologies from multiple domains and producing customized results depending on the user. A strategic approach that incorporates interdomain ontology search by means of computation of concept similarity has been proposed. The approach focuses on a key strategy of semantic set expansion where a concept similarity-based set inclusion has been focused using several variants of inter-domain relevant ontologies. A query expansion based on SPARQL based querying using generated keywords have been time. Furthermore, Precision, Recall, Accuracy and F-Measure has been increased for the proposed approach.

Organization: The remaining of the paper is organized as follows. Section II describes the related works. An overview of problem definition is given in Section III. Section IV depicts the Proposed Architecture. Section V demonstrates the implementation and the empirical results of our system. Finally, the paper is concluded in Section VI.

II. RELATED WORK

Awny Sayed et al., [1] has proposed an Arabic search engine named CASOnto. The proposed methodology performs searching in two types, namely, the keyword-based search and the semantics-based search in both the languages of Arabic and English. Resource description framework data and ontological graph are used to build the engine. Cristiano Rocha et al., [2] put forward an architecture for the search engine that considers a combination of classical search techniques, along with activation techniques for spread, applied to a domain-based semantic model. R. Guha et al., [3] proposed a framework and convenient application of the semantic search which is basically designed to improve the web searching experience. The paper also describes the implementation of two semantic search systems, which interlink classical search results with the relevant data collected from a huge number of distributed and random sources.

Vignesh Sivakumar et al., [4] put forward a semantic meta search engine which uses the technique of semantic similarity measure, which is used to refine the input query in a highly specific way. The model starts off by the input of a query by the user to Wordnet ontology, the subsequent process involves obtaining the neighbouring keywords followed by the selection of the most suitable query words by the semantic similarity between the query and the neighbours. Then, the ranking measure is used to rank the pages obtained as an output. Heiko Dietze et al., [5] proposes GoWeb, a combination of traditional keyword-based web search, including text-mining and ontologies to navigate large result sets and facilitate question answering.

In [6-7] an Ontology based Semantic Web search engine has been proposed. The models incorporate relationships between the keywords at all the relevant locations, exploring the user's personal interest. Ontology is used to calculate the similarity of these relations on each page, which will help us determine their relevance between the query provided and the user. Axel J. Soto et al., [8] proposed Thalia, a semantic search engine that has the capability to recognize eight different types of concepts that frequently occur in biomedical abstracts. The aforesaid search engine is updated daily from PubMed.

B. Fazinga et al., [9] proposes an approach called the FGGL Approach that is based on a structured query language, allowing the formation of complex search queries that are not ontology-based. The data on the web is carefully mapped onto an ontological Knowledge Base, which allows semantic search on the web relating to the underlying ontology. Wenwen Li et al., [10] proposed a semantic search engine that aims to spatial-aware search more intellectual and intelligent. The needed data is accumulated from a variety of geospatial resources. It uses an ontology-based information model to structure its unstructured data, specifically; it is based on SWEET (Semantic Web for Earth and Environmental Terminology) ontology. In [11-20], authors have proposed ontology-based search engines or other applications, which definitely contributes to the semantic web.

III. PROBLEM DEFINITION

Search engines provide a solution to users' queries by evaluating them based on their face values, that is, the meanings of the words. Multiple issues like lack of enough domain expertise and poor understanding of the correlation of similar terms lead to undesirable results arising out of a search engine. A semantic search engine considers the user history as well as background knowledge from the search activities of other users, to come up with a unique list of recommendations for the search queries. This is achieved with the help of implementation of multiple ontologies, corresponding to respective domains and designing cross-correlation between them.

IV. PROPOSED ARCHITECTURE

The architecture of the proposed system incorporating the Web-based Semantic Search Engine is depicted in Fig.1. It shows the important modules of the system, which are divided further into three individual phases. For Query Expansion, instances derived from SPARQL, have been used to query the ontology and returns an expanded set of query words. A limit is set to restrict the size of the expanded set. The subsequent process is Query word Generation for which expanded query is taken as input and is tokenized by removing unnecessary words thus preparing it for further recommendation algorithms, in which the function of recommending the most relevant URLs by semantically matching the query words to the dataset is done. Different weights are assigned which are used to calculate the final similarity score.

Next comes computation of Concept Similarity in which semantic similarity between two set of sentences is performed which is further used to rank the URLs in the

dataset. Finally, in the process of Personalization, the order of recommendation is shuffled based on user history and preferences. The function of updating user history upon clicks is also done.

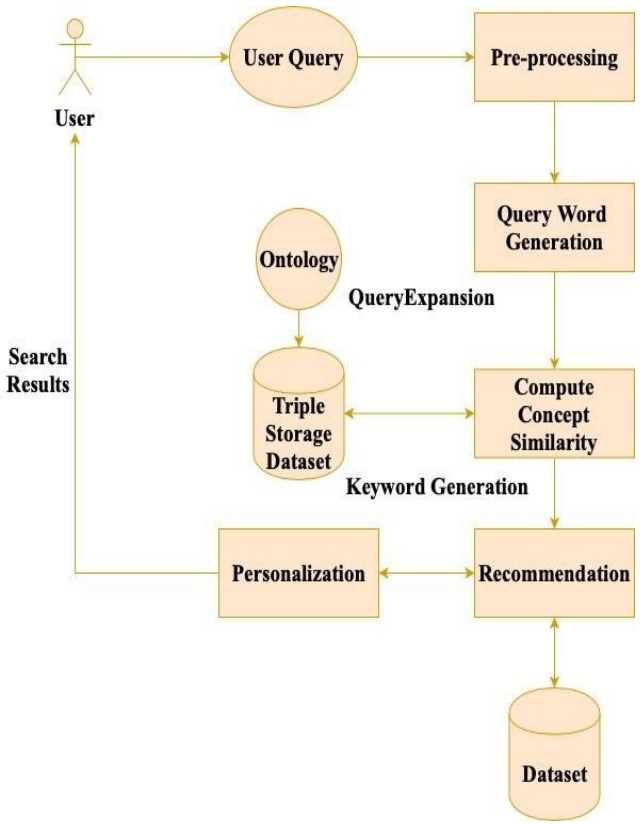

Fig.1: Proposed Architecture

Phase 1 mainly concentrates on the creation of ontology on multiple domains and storing them efficiently. The reason for facilitating modelling of ontologies as a separate phase is because various ontologies can be added to the existing ontologies whenever required. The ontologies are stored in a Triple store database using Apache Jena-Fuseki to retrieve the data faster and more efficiently. Phase 2 deals with the user interface, query pre-processing and semantic set expansion. The semantic search engine provides a user interface that enables the users to interact easily by creating user profiles for the users. The user interface enables auto completion of the query and also accepts misspelt words in the query and gives relevant output. Query pre-processing is an important task of every search engine to retrieve optimal result faster. It includes query extraction, pre-processing, indexing, ranking, etc. Query set expansion is the process of querying the ontology with the user query words using SPARQL and to augment the user's query set with additional related terms from the ontology in order to bring semantic into action and finally obtain the expanded query set.

Phase 3 mainly concentrates on Recommendation and Personalization of the search results. A dataset containing several URLs and corresponding tags is used to recommend relevant URLs based on the expanded query set using Semantic Similarity. After evaluating the semantic similarity of the expanded query set with the tags set of the dataset, the top URLs with highest similarity are selected to be returned as the result. Now Personalization comes into play, to make the search engine results customizable for the user, user profiles are used where the user logins with the user-name before the search is performed and then customizable search results appear i.e., different results appear for different users based on their priorities. This is achieved by recording each user's history and a count of no. of times that URL was visited. Finally, after undergoing Recommendation and Personalization, the search results appear in the user interface.

V. IMPLEMENTATION AND PERFORMANCE ANALYSIS

The implementation is majorly done using Python (version 3.7.3). The user interface is designed using JavaScript, PHP and CSS. XAMPP server is used to create a cross-platform web server for testing and execution. Python NLTK is used for pre-processing the input query. Protégé is used for Ontology modelling and as a knowledge management system. The ontology consists of three different domains, namely, Physics, Movies and Television shows in the OWL format, i.e., Ontology Web Language. The ontology is stored as Turtle Terse Language (TTL) - a data description language for the semantic web. Upon modelling, the ontologies are then stored as triples in a Triple Store Database using Apache Jena Fuseki. The triples are stored in a persistent dataset in the TDB which uses a custom implementation of threaded B+ trees. Once the ontologies are persistently stored in the Jena Fuseki server, integrated with the rest of the system by using SPARQL queries. The Jena Fuseki SPARQL server is queried using the Python RDF library. The Python Spacy library is used for calculating the semantic similarity between two words. The dataset contains URLs and their respective tags in a CSV format.

Table 1: Proposed Algorithm

Input: Query from the User
Output : Personalized search results

begin
Step 1. Input Query from the user (Q)
Step 2. Pre-process the text (i.e., tokenization and lemmatization)
Step 3. Save pre-processed text in the set (Q_w)
Step 4. c ← ConSim(Q_w.ele( ), Tonto.ele)
if(c > 0.5)
ROnto ← TOnto.ele( )
end
Step 5. for each (TOnto.ele( ))
c_1 ← ConSim (TOnto.ele( ), K_w)
if (c_1 > 0.75)
Re.K_w.URL( );
end
Step 6. Personalize using User's Historical Data
Step 7. Recommend the final search results
end

Upon pre-processing of the query words, the query set is expanded using the strategy adopted for Query Expansion as depicted in Table 1. It takes pre-processed query set and the ontology as input. The ontology is queried to obtain the subject, object and predicate of related words and SPARQL is used to return the expanded set of query words. Further, the scheme for keyword generation is used to generate a tokenized set of keywords by eliminating generic and stop words, by the incorporation of a Blank Space and Special Character Tokenizer and Lemmatizer. Further. the function of recommending the most relevant URLs by semantically matching the query words to the dataset is realized. Higher weights are assigned to the original query words. The URL entries are sorted as per their similarity scores. Entries with score greater than a threshold value alpha are considered for personalization.

Table 2. Evaluation of metrics for sample queries

	Physics queries	Movie queries	Inter-domain queries
Retrieved relevant	96	95	97
Retrieved irrelevant	4	5	3
Not retrieved relevant	3	2	3
Not retrieved irrelevant	0	0	0
% Precision	96	95	97
%Recall	97	98	97
%Accuracy	93.2	93.1	94.1
%F-measure	96.5	96.5	97

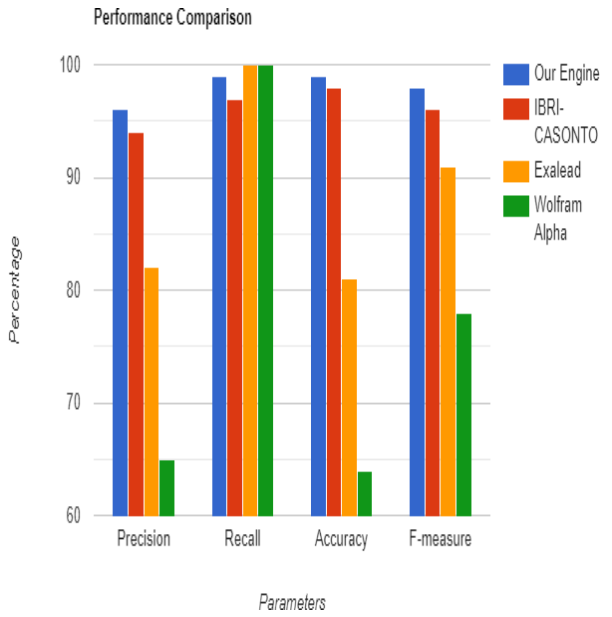

Fig 2. Performance Comparison of the Proposed Strategy with other existing strategies

Further, the semantic similarity between two set of sentences which is further used to rank the URLs in the dataset. Finally, the proposed framework takes into account the user history and preferences. The recommended URLs are further re-ordered by considering user history. Additionally, it dynamically facilitates updating the user history upon clicks. Personalization is done initially by storing search history for each and every user in which the visiting time-stamp, visited URL, count and other parameters are stored. Then this history is recorded and updated each time a URL is visited, be it new or old. After the Recommendation phase, i.e., after getting the result URLs, the parameters decide the precedence of the URLs in the result. The analysis evaluation of the search engine is measured on the basis of different metrics. This section evaluates, and analyses it based on the four important metrics: Precision, Recall, F-measure and Accuracy. Based on the combined results of all the queries sample, Precision, Recall, F-measure and Accuracy are evaluated for each domain as depicted in Table 2.

Based on the evaluation metric values obtained, the proposed model is put in comparison with various existing ones which include, IBRI-CASONTO, Wolfram Alpha, and Exalead as depicted in Fig. 2. The search engine-built retrieves more efficient results than the other engines, as the overall accuracy is higher. Thus, it is built according to the ontological-domain specific, highly scalable performance and handles both simple and complex queries by understanding the context behind the query. The proposed system yields an average precision of 96%, average recall of 97.3%, and average F-measure of 96.64%. The proposed search engine efficiently the needed information by making use of various mapping techniques that take place between classes and instances, due to which there is no need to go through a large number of results, that typically takes place in the case of keyword-based search engine. The approach proposed is highly efficient, scalable and can easily be modified to add many other domains. The system can also be modified to recommend various other domains apart from URLs as per the dataset used. The proposed design also includes the features of auto complete and spell check, like any usual search engine.

VI. CONCLUSIONS

With the constant evolution of the web, the problems of obtaining irrelevant and imprecise results will continue to bother the internet users. A strategic approach for interdomain web search using ontologies has been proposed. The proposed semantic search encompasses varied strategies like the semantic set expansion as well as concept similarity computation. The experimentations have been conducted for domain specific as well as interdomain queries and the exploratory semantic search scheme has been computed. Personalization has been achieved based on the web usage data of the users. The proposed strategy focuses on rendering multiple ontologies to find synonymous terms for query expansion. Expanded keywords are used to extend and improve the semantic interpretation of the input to be able to include more relevant results. The strategy proposed focuses on reducing the

irrelevance of search results and an overall F-Measure of 96.64% has been achieved.

REFERENCES

[1] Awny Sayed, and Amal Al Muqrishi. "CASONTO: An efficient and scalable Arabic semantic search engine based on a domain specific ontology and question answering", International Journal of Web Information Systems, Vol. 12 Issue: 2, pp.242-262, https://doi.org/10.1108/IJWIS-12-2015-0047, (2016).

[2] Rocha, Cristiano & Schwabe, Daniel & Poggi, and Marcus. A hybrid approach for searching in the semantic web. Thirteenth International World Wide Web Conference Proceedings, WWW2004. 374-383. 10.1145/988672.988723, 2004.

[3] R.V. Guha, R. McCool, E.Miller (2003) "Semantic Search", Proc. 12th International World Wide Web Conference (WWW '03), pp. 700-709

[4] Vignesh Sivakumar. Semantic Meta Search Engine using Semantic Similarity Measure. International Journal of Information System and Engineering Vol. 3 (No.2). 0.24924/ijise/ 2015.11/v3.iss2/73.79, 2015.

[5] Dietze, Heiko & Schroeder, Michael. (2009). GoWeb: A semantic search engine for the life science web. BMC bioinformatics.

[6] Sangeetha, A & Chidambaram, Nalini. (2018). Personalized semantic web search using ontology construction. International Journal of Pure and Applied Mathematics. 119. 16313-16320.

[7] Yong, Qi & Peijie, Hao & Yu, Hou & Ya-nan, Qiao. (2007). The Research and Design of the Semantic Search Engine Based on Ontology. 610-611. 10.1109/SKG.2007.274.

[8] J Soto, Axel & Przybyła, Piotr & Ananiadou, Sophia. (2018). Thalia: Semantic search engine for biomedical abstracts. Bioinformatics(Oxford,England).10.1093/bioinformatics/bty871.

[9] Fazzinga, Bettina & Lukasiewicz, Thomas. (2010). Semantic search on the Web. Semantic Web. 1. 89-96. 10.3233/SW-2010-0023.

[10] Li, Wenwen & Yang, Chaowei. (2008). A Semantic Search Engine for Spatial Web Portals. II-1278. 10.1109/IGARSS.2008.4779236.

[11] Deepak, Gerard, and Sheeba Priyadarshini. "A hybrid framework for social tag recommendation using context driven social information." International Journal of Social Computing and Cyber-Physical Systems 1, no. 4 (2016): 312-325.

[12] [Gulzar, Zameer, A. Anny Leema, and Gerard Deepak. "Pcrs: Personalized course recommender system based on hybrid approach." Procedia Computer Science 125 (2018): 518-524.

[13] Deepak, G., and Z. Gulzar. "Ontoepds: Enhanced and personalized differential semantic algorithm incorporating ontology driven query enrichment." Journal of Advanced Research in Dynamical and Control Systems 9, no. Special (2017): 567-582.

[14] Deepak, Gerard, and Dheera Kasaraneni. "OntoCommerce: an ontology focused semantic framework for personalised product recommendation for user targeted e-commerce." International Journal of Computer Aided Engineering and Technology 11, no. 4/5 (2019): 449-466.

[15] Chen, Mengxiang, Beixiong Liu, Desheng Zeng, and Wei Gao. "A Framework for Ontology-Driven Similarity Measuring Using Vector Learning Tricks." Engineering Letters 27, no. 3 (2019).

[16] Giri, G.L., Deepak, G., Manjula, S.H. and Venugopal, K.R., 2018. OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation. ICCIDE 2017, p.265.

[17] Pushpa, C. N., Gerard Deepak, J. Thriveni, and K. R. Venugopal. "A Hybridized Framework for Ontology Modeling incorporating Latent Semantic Analysis and Content based Filtering." International Journal of Computer Applications 150, no. 11 (2016).

[18] Giri, G. Leena, Gerard Deepak, S. H. Manjula, and K. R. Venugopal. "OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation." In Proceedings of International Conference on Computational Intelligence and Data Engineering, pp. 265-275. Springer, Singapore, 2018.

[19] Deepak, Gerard, Ansaf Ahmed, and B. Skanda. "An intelligent inventive system for personalised webpage recommendation based on ontology semantics." International Journal of Intelligent Systems Technologies and Applications 18, no. 1/2 (2019): 115-132.

[20] Deepak, Gerard, J. Sheeba Priyadarshini, and MS Hareesh Babu. "A differential semantic algorithm for query relevant web page recommendation." In 2016 IEEE International Conference on Advances in Computer Applications (ICACA), pp. 44-49. IEEE, 2016.