DOI 10.29042/2018-3927-3931

Helix Vol. 8(5): 3927- 3931

Semantic Information Search Based on Graph Analysis

*[1]Kalyani M. Dakhare, [2]Avinash J. Agrawal

[1,2]Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India

Email: kalyani1500@gmail.com

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract

Purpose of this paper is to resolve the problem of traditional search engines which hardly provide the essential content relevant to the user query. Current techniques for processing of query are mostly based on keywords. Thus, they have limited capabilities to understand the concepts and meaning involved in the user query. The solution for this is a semantic information search. A semantic search overcomes the drawbacks of mismatch associated with traditional keyword-based search. We create a semantic information retrieval system by combining Natural Language Processing and Graph Analysis. The idea of higher-level conceptual understanding of queries is developed to overcome the limitation of the traditional model. We present the method for semantic information retrieval by creating a semantic graph of the query. We then add synonyms and hyponyms of the query terms present in the semantic graph to retrieve the accurate documents. The level of accuracy will be enhanced since the query is analyzed semantically.

Keywords: Graph Analysis, Information Retrieval, Natural Language Processing, Semantic Graph, Semantic Search

Introduction

Current query processing techniques are depends on keywords. Thus, it has very few ways to grasp the concepts present in the user query. The higher-level conceptual understanding of queries is the main idea of this paper. As traditional information retrieval system is mainly based on keywords thus it is simple and easy. However, traditional system is based on the simple syntax matching with lack of understanding, leading to unsatisfactory search quality and results. Even though the query processing system has various enhancements, but have various limitations in terms of searching as it searches based on literal things. To overcome the limitations of the traditional model i.e. keyword-based model, we introduce a semantic search [1, 2] for information retrieval.

A semantic search overcomes the limitation of conceptualization associated with keyword-based search. The traditional information retrieval system based only on the occurrence of words in a document and ignores all other words associated with the keyword. For example, synonyms, hyponyms, etc. For semantic search, we used the graph-based method. The graph-based method used concept hierarchy and also supports the logical inference of the query. Semantic information retrieval has become an important part of IR field. The natural language processing (NLP) [3] method gives us the ability to represent keywords in terms of different part of speech. Using this we can find the relationship between various terms in the user query. By creating the semantic graph we can find the level of the graph and give weights to them for categorizing important terms. Addition of synonym and hyponyms finds more accurate documents. In the keyword based system, synonym and hyponyms are not taken into consideration thus some related document is not retrieved. The proposed method gives us the ability to semantically retrieve the relevant document.

Literature Review

Semantic IR has become the most important part of any search engines. Many papers describe the methods for this having various limitations. Our view of the semantic retrieval problem is very close to an ontology-based approach. In this approach [4] the user query is expressed in SPARQL which is an ontology-based query language. For indexing and processing of query, the external resources are used. An ontology-based information retrieval approach used inverted index which contains semantic entities associated with the documents. Heterogeneity, usability and scalability are some limitations of this approach on the web. Another method is Vector space model [5] which considers numerical feature vectors in a Euclidean space. The limitations of this model are the meaning of a text and structure cannot express, if two documents have similar meaning, but they are of different words then this model cannot represent word appearance sequence and relations, then similarity cannot compute easily. Next system is hybrid semantic search engine [6]. Hybrid semantic search engine is a combination of semantic search and traditional text search. Another semantic search method for the semantic web [6] is presented by author Ning et al. which describes a ranking algorithm and search algorithm. Representation of data is focused on a weighted directed graph. RDF tuple is developed based on the weighted directed graph. The ranked objects can be browse by random surfer. For the purpose of querying objects in XML documents, many search engines are adopted. The categorization XML search by the author Luk et al. is as follows, 1.full-text search, 2.to filter

information to discard, 3.an XML assisted search and 4.use of XML to translate queries among database [7].

A traditional search engine fails to provide the most relevant content for a query. To get the most relevant and useful content or documents for a user query, the query needs to be processed semantically along with keywords. The keywords in a user query could also have synonyms or a specific meaning semantically, and would be very useful to get the most accurate search results for a query within less time. Thus we have proposed semantic search system based on graph analysis.

Design Methodology

The proposed approach is explained in the block diagram below:

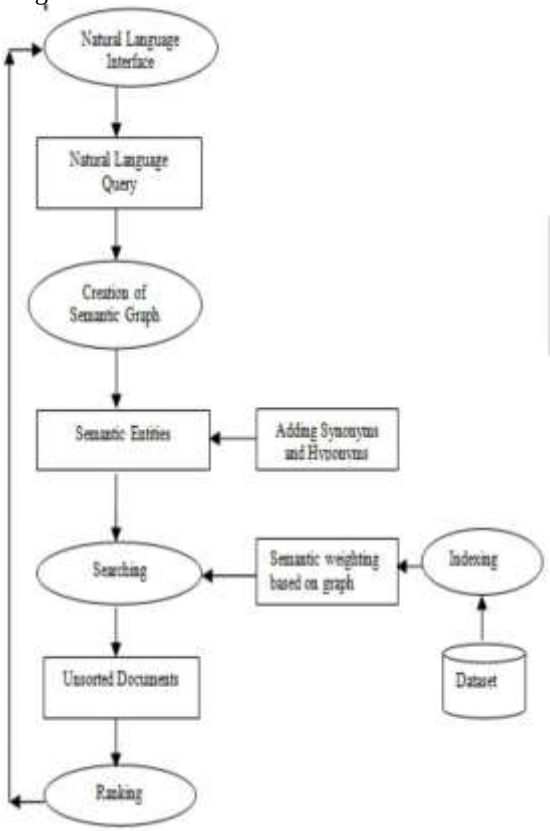

Figure 1.Block Diagram of the Proposed System

In the traditional semantic model, the user query is expressed in language like SPARQL which is an ontology-based query language and also contains SPARQL editor, where a user needs to have prior knowledge of that language. In the proposed system, natural language processing overcomes the limitation of usability of the traditional semantic model. By using natural language interface [8], the system becomes user-friendly and easy to handle by anyone. For research point of view, we used graph creation method [9] in a proposed system.

Algorithm:

Step1: Accept input from the user in terms of natural language query through natural language interface.

Step2: Create a semantic graph from user query.

Step3: Find synonyms and hyponyms of semantic entities if any, present within the semantic graph.

Step4: Create an index using semantic entities and their respective synonyms & hyponyms and assign level wise weights**.**

Step5: Search the documents in the datasets using semantic graph created in step2 and rank the documents accordingly. Display the matched documents to user according to rank.

Step6: Comparative analysis of accuracy. Compare proposed approach with the keyword-based search.

Module Description:

Generation of the semantic graph from query

Using graph creation method we represent user query in terms of the semantic graph [10, 11, 12]. This is a preprocessing module of the proposed system. This module has three steps, Input query:

The system accepts user-entered query as an input using natural language interface. The user-entered query is then sent for further preprocessing.

Tokenization:

In the process of tokenization, the user query gets divided into a number of different tokens i.e. each and every word of query gets separated. The process of tokenization is important to achieve tagging.

POS tagger:

The part of speech tagger (POS tagger) tags each and every token obtained from the process of tokenization. POS tagging is done with different part of speech like a noun, adjective, pronoun, etc. The system can removes stop words and create the semantic graph by using POS tagger. To create a semantic graph, we then extract the entity and attributes from the query where a noun is considered as an entity and adjective as an attribute. By considering domain at the top level, entities at the second level and attribute at the last level, a semantic graph of user entered queries is designed.

Adding Synonyms and Hyponyms

Various constituents of the input query like a noun phrase, adjectival phrase, etc. are obtained from the output of preprocessing module. This output which is in the form of semantic entities is then checked for the synonyms and hyponyms of each term. In all traditional IR systems, the terms that are actually present in the query are only considered for searching purposes. But many terms may have synonyms which have the same meaning of that term or they may have hyponyms which express that term effectively and so no. The traditional IR system does not use such terms, resulting in less accurate results of query search. Thus, to overcome this drawback, in the proposed system we added the synonyms and hyponyms of the terms present in the semantic graph, which leads to retrieving accurate documents for the user query. The synonyms and hyponyms are fetched from the specially created data files for

synonyms and hyponyms. Thus, in a proposed system along with the keyword in a query, their synonyms and hyponyms are also used.

Searching

In this module, the system assigns weight to all levels of the semantic graph for quick and fast search. The weight assignment is done according to levels of semantic graph. The highest weight is assigned to a top level of graph and lowest weight assigns to the bottom level of the graph. The system can find the most important and less important term within the query by adding the level wise weights to the semantic graph which will result in retrieval of more accurate and related documents.

Weights assignment: To find the importance of the particular term in the query, the system assigns weights to each and every term in the semantic graph. We assign weight to terms in a semantic graph with the help of a predefined set of values. Table1 shows the level wise assignment of weights:

Level	Term	Synonym	Hyponym
Top	1	1	0.75
Middle	0.75	0.75	0.5
Bottom	0.5	0.5	0.25

Table 1.Weight Assignment

The weighted terms along with synonyms and hyponyms, if any are then arranged in the index. To find the score of each document, we used a TF -IDF algorithm. The semantic TF -IDF score is the product of assigned weights of each term and their respective TF-IDF score. By using this score we can able to show the difference between the normal TF-IDF score and semantic TF-IDF score.

Ranking

Using all modules described above, the proposed system retrieved all relevant documents of user entered query. The ranking is the process of arranging these retrieved documents into the sequence. We ranked relevant documents according to their score, where the highest scored document is present at first position then second highest and so on. We ranked documents differently using following methods:

Manual ranking:

For creating a benchmark for the proposed system we performed manual ranking. To rank documents accordingly, we created set of queries and considered one dataset. By understanding and reading dataset thoroughly, we performed manual ranking for a set of queries. This benchmark is used to find the accuracy of different IR systems.

Ranking using normal TF-IDF score:

The system in the previous module, calculates the normal TF-IDF [13] score of all documents presents within dataset according to the user entered a query. We ranked them according to calculated score. We used this ranking to calculate the accuracy of normal TF-IDF system by comparing it with manual ranking.

Ranking using semantic TF-IDF score:

By using semantic TF-IDF score calculated in the previous module, we ranked documents within the dataset. This semantic ranking is also compared with manual ranking for calculation of accuracy of semantic TF-IDF scoring.

Results and Discussion

To test the proposed work, a set of NL queries for the mobile domain are prepared. Consider the example below to understand the proposed system.

Example: User entered query:-

"Good Camera and Best Music Phone"

The entered query is sent to the preprocessing module where system creates the semantic graph.

Constraints of the graph:

Domain-Phone

Entity-Camera, Music

Attribute-Good, Best

To form a graph, system mapped constraints of the graph together by considering noun as an entity and adjective as an attribute. The relationship between entity and attribute is,

Camera – > Good

Music – > Best

The graphical representation of user entered query i.e. a semantic graph is shown in Figure 2. This semantic graph shows the meaning of user entered query.

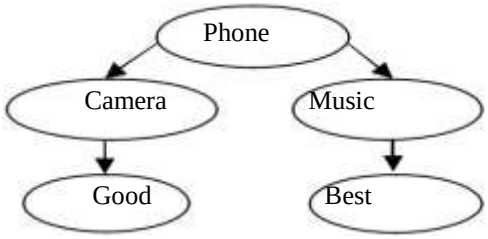

Figure 2. Representation of Semantic Graph

By using keywords present in the semantic graph, the synonyms and hyponyms of respective terms if any, are fetched to get more relevant and accurate document. The index is formed by adding both keywords and their respective synonyms and hyponyms if any. As shown in table1, the weights will assign to each keyword and their respective synonym and hyponym in the graph based on the levels of the graph. For a given query, the weighted index is, Phone (1), Mobile (1), Handset (1), Camera (0.75), Music (0.75), Opera (0.5), Good (0.5), Best (0.5).

The index term phone is present at top of the graph and has two synonyms, mobile and handset. Thus, the system assigns the highest weight to term phone i.e. (1), and because it having the synonyms, the same weights are assigns to them. The term camera is present at the middle level, i.e. it is an entity. Thus we assign a weight (0.75) to the term camera. Index term music is present at the middle level, thus weight

is (0.75) and also have hyponym opera thus according to table1, the system assigns the weight to this hyponym as (0.5). Good and best is present at a bottom level of semantic graph thus, the system assigns the lowest weight to them as (0.5) as they are attributes of the graph.

Using this index, the system searches for the documents related to query in the dataset using TF-IDF algorithm. As it contains the synonyms and hyponyms along with the key terms, the result obtained is more accurate than normal TF -IDF algorithm. The next step of the system is to rank the retrieved documents. The ranking is performed on the basis of TF-IDF score of each document. The semantic TF-IDF score is calculated by multiplying TF -IDF score with applied weights at each level of semantic graph. Based on this semantic score, the ranking of the document takes place in which document with the highest score is placed at the first position and so on.

When the proposed system was tested, a marked improvement in accuracy was observed for most of the queries. The table2 shows the accuracy of our system comparing it with the traditional keyword-based system, for few sample queries. The table below depicts that the accuracy value of our system is higher than the values obtained in the keyword-based search. Accuracy is computed as a percent of documents matching with the benchmark ranks we computed using the manual rankings as stated above, by understanding and reading the test dataset thoroughly, and ranking the documents for a set of queries with the order of most relevant and useful content.

Sample Queries	Semantic search	Keyword -based search
Good camera and best music phone	78%	49%
Phone with high-resolution camera	72%	30%
Top mobile company with high reputation	81%	54%

Table 2.The Accuracy of Sample Queries

When the conventional and proposed systems were tested against the benchmarks, a marked improvement in accuracy was observed for most of the queries using the proposed system. Figure 3 shows the accuracy of semantic search Vs traditional keyword-based search. The graph is drawn taking the 25 queries into consideration, their corresponding accuracy is plotted for both proposed and traditional systems. The curve in the graph clearly depicts that semantic search system i.e. proposed system retrieves the accurate links for the user query.

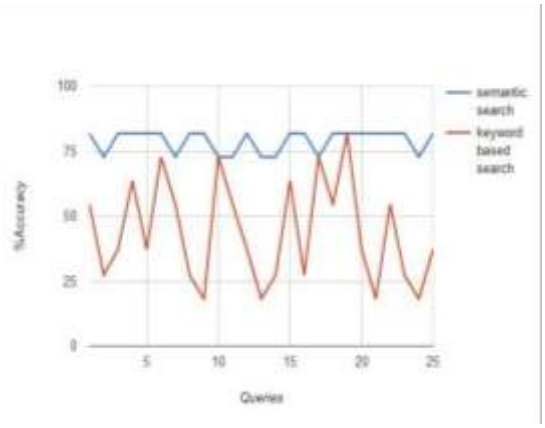

Figure 3.Accuracy Vs Query Graph for Semantic and Keyword-Based System

Table3 shows the average accuracy of semantic search and keyword-based search. From this, we can infer that the higher value of average accuracy for our proposed system i.e. semantic search system when compared to keyword-based search system depicts that our system has a better accuracy and performance in retrieving the relevant results than the keyword-based search systems.

	Semantic Search	Keyword-based Search
Average Accuracy	77%	51%

Table 3.Average Accuracy of Semantic and Keyword-Based System

The higher value denotes the best coverage of our system compared to the traditional keyword-based systems.

Conclusion

In this paper, we have presented a semantic information retrieval model which, extends the traditional information retrieval model, addresses the challenge of conceptualization, and integrates the advantages of both semantic-based and keyword-based information retrieval. By understanding contextual meaning and searcher intent, semantic search improves search accuracy. For the various domains of information search like mobile, vehicle, books, medical, government information systems, etc., the proposed system will be useful. The result of evaluation has shown that the semantic information retrieval model performs much better than traditional keyword-based information retrieval model. The accuracy and retrieval of the relevant document are improved in the semantic model. Future research will aim to present the higher level of semantic search, like one can run this system for World Wide Web to access relevant documents by adding various domains into the system, with the last aim to provide higher level of query understanding when answering user's needs and retrieve accurate documents.

References

1. A.Tulika Narang, B.Prof. R.R. Tewari, “Towards Semantically Enhanced Information Retrieval”, International Journal of Latest Trends in Engineering and Technology (IJLTET), 2012.
2. H. M. Harb, M. F. Khaled, M. N. Nagdy, "Semantic Retrieval Approach for Web Documents", *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 2, no. 9, pp. 11-75, 2011.
3. Nadia Soudani, Ibrahim Bounhas, Yahya Slimani, “Semantic Information Retrieval: A Comparative Experimental Study of NLP Tools and Language Resources for Arabic”, Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on 6-8 Nov. 2016.
4. Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, Enrico Motta, "Semantically enhanced Information Retrieval: An ontology-based approach", Departamento de Ingeniería Informática, Universidad Autónoma de Madrid, Madrid, Spain, 2011.
5. Daiss, Jae-Yong Chang and Il-Min Kim, "Analysis and Evaluation of Current Graph-Based Text Mining Researchers", Advanced Science and Technology Letters Vol.42, 2013.
6. Hai Dong, Farookh Khadeer Hussain, Elizabeth Chang, “A Survey in Semantic Search Technologies”,IEEEXploreDOI:1109/DEST.2008. 4635202, March 2008.
7. R. W. P. Luk, H. V. Leong, T. S. Dillon, and A. T. S. Chan, "A survey in indexing and searching XML documents", Journal of the American Society for Information Science and Technology, vol. 53, 2002, pp.415 - 437.
8. Kalyani M. Dakhare and Dr. Avinash J. Agrawal, “Optimizing Information Search using Graph Analysis”, International Conference on Recent Trends in Engineering & Sciences (ICRTES), February 21,2018. [Presented]. To be published in coming issues Vol 8 of International Journal of Engineering & Technology (UAE) ISSN:2227-524X [SCOPUS Indexed].
9. S. S. Sonawane, Dr. P. A., “Graph-based Representation, and Analysis of Text Document: A Survey of Techniques”, International Journal of Computer Applications, 2014, Volume 96
10. Wei Wei Jin and Rohini Srihari, “Graph-based text representation and knowledge discovery’, In Proceedings of the SAC conference, 2007
11. Jure Leskovec1, Marko Grobelnik2, and Natasa Milic-Frayling3, “Learning Semantic Graph Mapping for Document Summarization”, International Journal of Web & Semantic Technology (IJWesT),2012.
12. Dr. Avinash J. Agrawal, Dr. O. G. Kakde, “Semantic Analysis of Natural Language Queries Using Domain Ontology for Information Access from Database”, I.J. Intelligent Systems and Applications, 2013, 12, 81-90
13. Mingyong and Liunand Jiangang Yang, "An improvement of TFIDF weighting in text categorization", International Conference on Computer Technology and Science (ICCTS 2012).

