Indian Journal of Science and Technology, Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/108025, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Integrated Search System using Semantic Analysis

M. Sujatha, T. Renugadevi, R. Ezhilarasie and A. Umamakeswari*

School of Computing, SASTRA University, Thanjavur - 613401, Tamil Nadu, India;
sujatha@cse.sastra.edu, renugadevi@cse.sastra.edu, ezhil@cse.sastra.edu, aum@cse.sastra.edu4

Abstract

Objectives: With the rapid growth of technology drastically there is an increase in users of computer system. Search system plays a vibrant role in optimizing the search time. **Methods/Statistical Analysis:** The integrated search system has the need of improving search accuracy and expanding their coverage of search. This can be done using concepts of semantic analysis. **Findings:** In this paper an integrated system is formulated by accumulating multiple sources such as local storage, secondary storage and online repositories. The keyword is contextually interpreted by making use of ontologies, using this interpretation of the keyword the required multimedia or/and textual data is searched as intended by the user. **Applications:** This integrated search system analyzes the meaning of the query and provides the search result according to the intention of the user through is proper expansion of keyword.

Keywords: Integrated Search, Ontology, Semantic Analysis

1. Introduction

In recent days, the search engine technology is the major concern for the developers. It created a revolution in bringing up people to use the Internet and various data sources to the optimum level. Various researches have been performed and tested for the accuracy of the search engines. It is not an exaggeration to say that it requires the same attention as many other applications that are fundamentally used. The search engines that are present today are efficient but not to the most heuristic measure. They are morphed into various forms and concentrate only on a particular data source. The other main striking issue, which is prevalent today, is that, the search mechanisms are done only based on the keyword that is being searched for. The concept of semantic analysis is being introduced to prevent irrelevant data to be displayed as the meaning of the keyword/query is searched. Considering this issue, ontology is constructed in this work, which diagrammatically represents the possible interrelated concepts within a domain.

Having search systems only for a data source in particular is quite cumbersome for the user who has a vague idea of where the data might be located. In order to remove this ambiguity this system proposes a search mechanism, which integrates many data sources and provides the user with the most appropriate result. Our proposed system does the aforementioned functionalities which is not present in the current search engines. In[1] authors proposed a search system which integrates search results from local storage and global repositories in a Smart TV. The system involves two phases which consist of interpreting the meaning of the user keyword using knowledge based server and the actual searching of the data from the storage repositories. Ontology is built for film domain and semantic analysis is done by knowledge base server using the techniques of stemming, term mapping and constructing and ranking of query graphs. The interpreted user query is directed to three different search engines Local repository search engine, EPG (Electronic Programme Guide) repository search engine and Global repository search engine. In[2] authors dealt about how the words are related with respect to their semantic property or meaning. They measured the similarity between the words using semantic relationships. The relationships between words with the same meaning are extracted and the document is automatically organized into meaningful categories to make search of data easier. The input text

*Author for correspondence

is analyzed using Natural Language Processing for deriving the meaning of the text and the correct meaning is selected from words having multiple meanings using Word-Sense Disambiguation. In[3] puts forth two approaches for Semanticaly annotating the images. Bottom up approach is followed to bridge the gap in semantic nature of words. The collection of words with similar meaning are grouped as semantic space. The words having the same meanings share the same positions in the semantic space. A series of observations are made on the image and terms/words are assigned to each image based on the image content. Two matrices are built, the first matrix represents the co-ordinates of the terms in the semantic space and the other matrix represents the co-ordinates of the documents in the semantic space. Top down approach is done by using ontology to describe and get the images. OWL (Web Ontology Language) is used in the RDF (Resource Description Framework) to build the ontology of images to increase search accuracy. Similarly images can be semantically searched. In[4] based on the attributes of the input query image in QBIC system. In[5] authors proposed an approach for video search, that splits the video into keyframes (video frames). The content of the video is observed and the keyframes are split based on the distinct movements in the video. In[6] authors elucidate that there is a need of efficient mechanisms to go through the content of the documents that are present in the World Wide Web and provide them indexes in such a way as to improve search efficiency. WebSeek is a method that is used to extract the visual information from the web and categorize these ideal information according to their themes. The WebSeek programs make use of colour histograms to perform content-based search. The procedure consists of three distinct steps. The web documents which contain images, videos and/or hyperlinks are found by the Web Agents, the URL of the images and videos are extracted and the information about the multimedia data are retrieved and categorization of content is performed. In[7] discusses about Latent Semantic Indexing. It is the process of retrieval of data from repositories using matrix computations. The commonly used retrieval technique poses ambiguity where many keywords could refer to the same information. This method is prone to failure. The matrix computation which is stated by them is called Singular Value Decomposition. This process filters out the ambiguity where more than one document has the same semantic meanings. This recurring property of the document is called noise in the document. Every image or picture is represented as a collection of pixels. The user does not refer to the individual pixels while searching but to an object as a whole. Keyword based search systems have the drawback that minor changes in the keyword would cause errors in results. Hence the Latent Semantic Indexing is useful in such cases to search using the semantics of the image. In[8] explains that the concept of content-based video object segmentation enables unprecedented functionalities in accessing contents and retrieval. In[9] author developed WebSeer, a process used to locate images on the World Wide Web. It indexes the images by making use of the content present in the image and also the associated text along with the particular image. The search is made more accurate using WebSeer because it classifies the images into half body, full body, etc. The recurring, repetition and monotonous images are found and omitted from the search. Color Histograms are used to find the similarity between different images in the complete database. The HTML documents are scanned to verify whether any other relevant information is contained in the documents. If so, that information is added to the texts associated with the image. WebSeer has the ability to classify photographs and artificial images. In[10] describes about how the search for images gets very easy when the concepts of semantic analysis is used. Text is used in terms of Semantic keywords to represent the image. From the perspective of the user, the requested query may be in many forms, which are, the parts of the image, information about the object and image color, texture and shape. In[11] describe that the semantic social network called Moveek is dealt with the developed annotation tool which is used to classify and process the words that are mentioned in the post that is being annotated by our semantic annotator. The web services which play an important part (as they tell the user how efficient the tool is) has been developed in .Net web services. The sites are developed using PHP and JavaScript languages. The web content is related using a specific knowledge representation which is referred to as semantic annotation. The GATE API is open source software which helps in the annotations. After the annotated terms have been formed and published in the website the users will be allowed to check publications by just clicking on the keywords the user wants in the query result. The architecture developed illustrates that one can consider the web services to do various tasks like information extraction in addition

to our semantic annotation tool which was developed. In Orchestration, all the interactions that are present in the business process are described in a traditional workflow system which is then executed by an orchestration engine. Collaborative nature is emphasised and centralization is reduced in Choreography. The messages which are considered as public are relevant and the services present only know about its own behavior and interactions. In[12] is dealt about how relationships are established from the unmanaged and structured data that are available in the social networking sites that are present today. All the details are converted to semantic format and are stored in the KDF file where the semantic searches can be done using the SPARQL queries. Extension of data within the RDF is done by OWL which also describes the relationship that exists among the data. Components which they used were RDF parser, RDF serializer, RDF store and RDF query engine. The RDF parser, which translates sequent character format into KDF triple format. The RDF serializer is used for the inverse function of the previous component. RDF store contains RDF triple format definitions and can also accommodate data from different sources. RDF query engine is used to retrieve data from RDF store. Conceptualization is a well-defined domain and is created accordingly. The relationships among the members in the domain are established accurately. Formalization is nothing but a language with clear semantic definitions and symbols to make user read easily. FOAF is an ontology that is used to describe the relationship between the user and the activities. Libraries such as Jena and ARC2 are used to integrate current databases to the web browser. User information is retrieved from the social networking sites and with the help of ARC2 the data gets converted into RDF format. This application enables the user to scout people who have the same interests. In our work a system is proposed for integrating multiple search system along with semantical analysis.

2. System Overview

The overview of the ASAP search system is shown in Figure 1. ASAP server acts as an interface between ASAP interface, ontology server and Data Repositories. ASAP interface is connected with ontology server and ASAP interfaces. User search queries are sent to ontology server and then routed to ASAP server after semantic analysis. ASAP server is responsible for coordinating with multiple data resources and extracting the data from them. It is also responsible for returning the results to ASAP interface. The architecture of ASAP search system is shown in Figure 2. It is broadly classified into four major constituents which are ASAP interface, Ontology server, ASAP server and Data Repositories. The analyzed meanings of the keyword are sent to the ASAP sever, which is the main component of the search system. It is the coordinating server, which coordinates the movement of data across the components of the system. The analyzed keyword is then distributed across the various data sources (local, secondary and online repositories). This keyword is then searched in the three repositories and the results are then sent back to the ASAP server, which supplies it to the user interface, which in turn gives the result to the user.

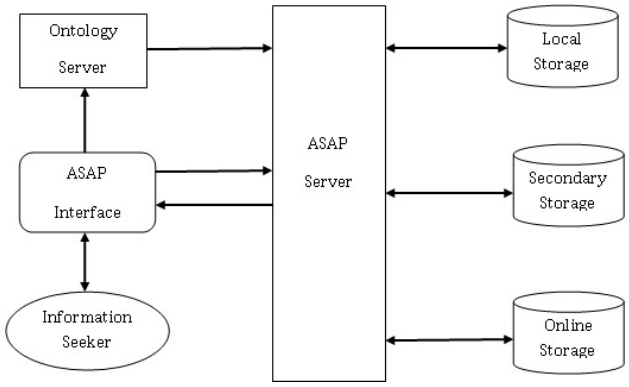

Figure 1. System overview.

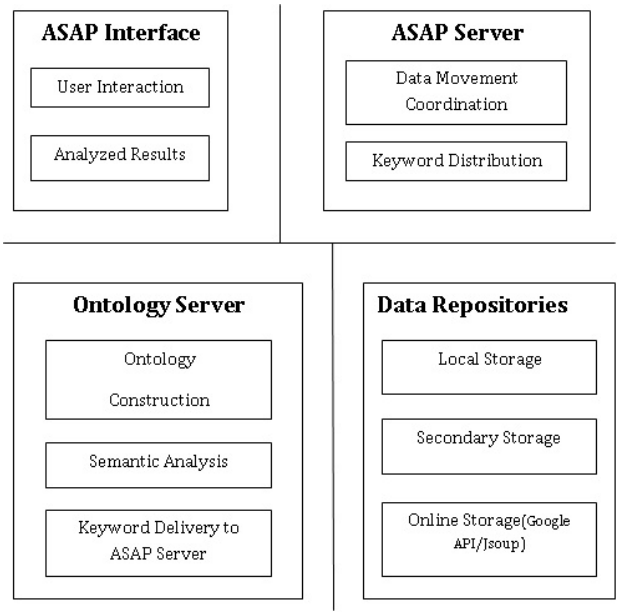

Figure 2. System architecture.

3. Architecture

3.1 ASAP Interface

This component carries two major tasks: User interaction and Displaying analyzed results. Information seekers interact with ASAP interface and provide the search keywords. The user given input keywords are received normally in the form of texts and forwarded to the ontology server. It also provides analyzed search result display, which is user friendly and easy to understand.

3.2 Ontology Server

The ontology server acts as knowledge base, maintaining the ontology of the data in a well-organized manner. Each and every class is created and the ontology is built based on relationship among the classes. Search keywords are mapped with classes residing in the ontology server and related objects are identified using their properties that have been crafted earlier. This makes the search for the meaning of the keyword and thus semantic analysis easier. Identified objects are further refined and dispatched to ASAP server.

Figure 3. Burger example.

3.3 ASAP Server

The ASAP server is the coordinating server, which serves as a central component of ASAP search system, used to transfer data from one component to another in the system. It is responsible for reception of the set of semantically analyzed keywords from the ontology server and distributing it to the various data sources for the actual search. Depending on the selected search category ASAP server will perform the search in the corresponding data resource. It is also responsible for receiving the searched results from the repositories. It then transports these results back to the ASAP interface component where the results are displayed in an organized manner.

3.4 Data Repositories

ASAP search system extracts the data from the repositories. This comprises of local storage, secondary storage and online storage. The data is searched for and the result is returned back to the ASAP server from where the results are sent back to the ASAP interface.

4. Implementation

An ontology server is built aided by the tool called Protégé. This tool builds the entire knowledge database in the most user friendly way. Ontology is created that brings the hierarchy levels into life that basically helps out in getting the best results. All the necessary classes and object properties that relate the classes with each other area created. The ontology is built in such a way that it builds a query graph with an ideology that the graph, which is the shortest, is the best result. There has been a concentration on few domains only as the abundance of domains is very huge. So, the food domain is taken as the primary consideration in the example. The ontology consists of various food items such as burger, pizza, dosa, naan, etc. which are represented as classes. They are related to each other through various relationships such as has Fillings, has Topping, is A base Of, etc. which are defined by the user. As shown in the Figure 3 we are taken the burger example where the classes which have been created are visible, for example we have the Angus burger class. Likewise all the classes are built with relations that can be semantically related in order to provide an analyzed result is shown in Figure 4. After the construction of the ontology, the ontology could be saved in a number of formats (e.g.: OWL, RDF, Turtle etc.). The major issue to be addressed is the connectivity of the knowledge database to a JAVA class. The ontology is saved in a Turtle format which has the various classes, object properties and relationships in a well organized way. The Turtle file is then parsed from the start to the end and the various classes, their object properties and relations are put into a separate text file. From this text file, the necessary keywords' classes and object properties

are identified based on the user input query and the corresponding words are returned back to the ASAP server. The phase 2 of the system begins from the ASAP server which co-ordinates the searches that are being done in the local, secondary and online repositories. For the local and secondary storage retrieval we use a simple JAVA search technique involving directories to get the respective data. These are also done using eclipse. To search from online repositories two methods were considered, which is: The use of the Google API, which is an interface in JAVA and the use of the methods of jsoup package in JAVA. Out of these two methods, the Google API was the more efficient and implemented. The final stages of the second phase and the system are etched out by integrating the results from all the three sources and an analyzed view of the data is portrayed to the information seeker. The system contains an ASAP interface. In our proposal the ASAP user interface is created with the help of eclipse a JAVA compiler.

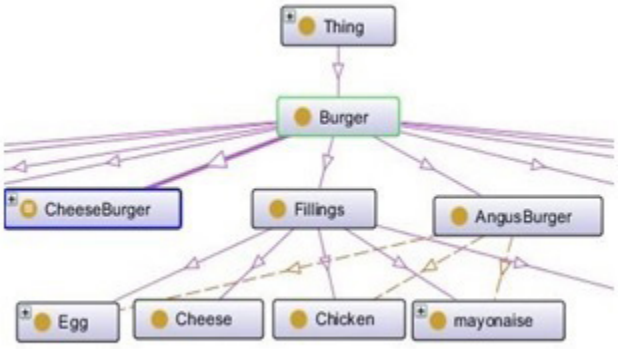

Figure 4. Class hierarchy.

5. Conclusion

The search systems today suffer from a lack of exact accuracy of the search keywords which are sent to the various data sources to be searched for. This implies that the search results which are returned are largely unwanted results which do not match the requirements of the user's needs. This reduces the efficiency of the search results. In this paper, we proposed the use of semantic analysis to understand the meaning of the keyword specified and then return search results for those set of words which is given the meaning of the user specified keyword. This minimizes unwanted results and improves the efficiency of the search by the use of ontology based semantic analysis. This semantic analysis is done by building the ontology of the chosen domain of data objects. These concepts can be further improved by the use of advanced semantic analysis using complex algorithms for large amount of real data.

6. References

1. Myung-Eun KIM, Joon-Myun CHO, Jeong-Ju YOO, Jin-Woo HONG, Sang-Ha KIM. A proposal of semantic analysis based multi-level search system for smart TV. ICACT Transactions on Advanced Communications Technology (TACT); 2013 Mar; 2(1):197–205.
2. Bollegala D, Matsuo Y, Ishizuka M. A web search engine-based approach to measure semantic similarity between words. IEEE Transactions on knowledge and data engineering. 2011 Jul; 23(7):977–990.
3. Bridging the semantic gap in multimedia information retrieval. Available from; http://eprints.soton.ac.uk/262737/
4. Flickner M, Sawhney H, Niblack W, Ashley J, Don BQ, Gorkani M, Hafner J, Lee D, Petkovic D, Steele D, Yanker P. Query by image and video content: The QBIC system. Journal of Computer. 1995 Sep; 28(9):23–32.
5. Hanjalic A, Lagendijk RL, Biemond J. A new method for key frame based video content representation. Image Databases and Multi-Media Search. 1998; 1–11.
6. Smith JR, Chaang S-F. Visually searching the web for content. Journal of Multi Media. 1997 Jul-Sep; 4(3):12–20.
7. On SVD-free latent semantic indexing for image retrieval for application in a hard industrial environment. Available from: http://ieeexplore.ieee.org/document/1290365/
8. Zhong D, Chang S-F. An integrated approach for content-based video object segmentation and retrieval. IEEE Transactions on Circuits and Systems for Video Technology. 1999 Dec; 9(8):1259–68.
9. Swain MJ, Frankel C, Athitsos V. WebSeer: An image search engine for the world wide web. Challenge of Image Retrieval; Newcastle. 1999. p. 1–8.
10. Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI. Matching words and pictures. Journal of Machine Learning Research. 2003 Jan; 3:1107–35.
11. Camarillo RP, Conde-Ram JC, Sanchez LA. A hybrid approach for solving the semantic annotation problem in semantic social networks. Research in Computing Science. 2013; 65:25–33.
12. Social Network Data Retrieving using Semantic Technology. Available from: http://ieeexplore.ieee.org/document/6605810/