International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 2 Issue: 2 08-Mar-2014,ISSN_NO: 2321-3337

Semantic Information Retrieval Model by Spectral Clustering

[1]Annie Jones, [2]Senduru srinivaslu

PG Scholar, Sathyabama University, Chennai,India [1]
Research Scholar, Sathyabama University, Chennai,India [2]
anniejonese@gmail.com, sendurusrinivas@gmail.com,

ABSTRACT— *The Web which is increasing day by day has huge volume of unstructured data, with several aims, qualities and aspects which makes retrieval a tedious task. Semantic web which extend our current web has a focus to retrieve the data more precisely with vocabularies. These vocabularies are understood by the people and computer. Ontology, the core concept of semantic web which explodes data from knowledge base consists of instance of classes. This paper proposes method on how retrieval of information semantically can be done from heterogeneous data store. Here clustering methodology is used to match both ontology and the user query. The algorithm will navigate into the deep roots of the ontology structure and group the similar nodes with the query. The collected similar data are stored in a buffer area to produce an optimized output. The clustering of the data is done semantically to achieve higher relevancy. The spectral clustering algorithm which is used to achieve clustering semantically will locate the sparsely located data and match them efficiently. The basic idea is to collate the web not only to link the large heterogeneous documents but also to instruct meaning of the information in those documents.*

Keywords— Semantic application, Heterogeneous, Ontology, Clustering.

1. INTRODUCTION

For years, the current web consists of interlinked documents that are on screened before people. WWW provided the documents and the humans have to connect the sources. These interpreting and connecting of sources are from unstructured data that are thrown before the users to understand and not by machines. Semantic web is working of computers cooperatively, to help people in finding relevant and required knowledge. The user will post the query and the machines will interact with each other to provide the relevant data. But in the current web, machines will not interact only the user has to interact. It's a smarter web that understands the meaning of documents. The information in the web is increasing day by day, so modeling the web is important and hard-hitting.

With the rapid development of computer network technology, the database technology and wide application of internet, a series of "Information Islands" came into being due to different running environments, different storage and representation of data. In order to make better use of these distributed, heterogeneous information resources, integration and sharing of heterogeneous data sources in the network becomes an effective way of eliminating the "Information Island". Heterogeneous data sharing is the premise of ensuring the integrity of data to provide users with a unified access interface. The different sources and different data formats in the logical or physical integration, packaging, handling, shielding the underlying data source differences. XML has a mature technology standard and improve the query mechanism, which fully meet the Internet and distributed heterogeneous environments, it can realize the data that the separation of content and data, enabling heterogeneous data sources through a unified data model the exchange of information, to some extent to achieve the data sharing, but the semantics of heterogeneous information sources is difficult to be effectively addressed. Artificial Intelligence, Ontology, Semantic Web technology for heterogeneous data integration provides new methods and ideas. At present, ontology has been described as a data source and the sharing tools applied to heterogeneous data. The three important concepts required are Conceptualization, Vocabulary, and Axiomatization. On the other hand, web based knowledge sharing activities demand that human and/or machine agents agree on common and explicit ontologies so as to exchange knowledge and fulfill collaboration goals. Semantic web is the data Integration at Web Scale. The web of data defines the framework for integrating multiple sources to draw new conclusions and architecture for describing all kinds of things, items, collections, services, processes, etc. They also increase the utility of information by connecting it to its definitions and its context. They provide effective management and reuse of data at various scales, personal, group, enterprise, community, web. Ontology has a similar philosophy and metadata definition: "the existence of the presence" in the knowledge sharing process, "ontology is a shared conceptual model of clear formal specification." Since the Semantic Web have been proposed, the ontology became interconnected content network resources to solve an important way, the semantic representation in the sharing of resources and reasoning play an important role. Currently, the establishment of various public shared ontology knowledge bases using ontology reasoning for knowledge reasoning research is under way. Heterogeneous data sharing, the ontology as a middleman and it shields the underlying structure of heterogeneous data sources. Users to perform queries, you cannot know the structure of various data sources, they do not know how to query data, only need to know what data the system provides a query for the body on it. System can be defined semantics and ontology mapping will automatically query is decomposed into a query for each. Semantic applications can be categorized on the characteristic it encompasses semantics in it.

2. RELATED RESEARCH

In tradition systems, the logic representation of user and the information provided are clearly based on the list of keywords (Guhas, McCool & Miller 2003). This will entail some information loss. This method of extracting is not expressive as it does not have any explicit relationship among the set of words. The information will be provided based on the list of keywords available in the user query.

Second type of system involves the natural language representation, (Lopez, Sabou, Uren &Motta, 2009) which extracts the syntactic information based on the linguistic analysis of the user data. The syntactic information is based on the subject, predicate and object of the sentence.

Controlled natural language (Bernstein & Kaufmann, 2006; Cohen, Mamou, Kanza &Sagiv, 2003; Gigunchiglia, Kharkevich & Zaihrayeu, 2009) the query is expressed in tags which facilitate easy processing and mapping of the properties and objects. These queries extend a hand in building the ontology or schema for the query providing space for semantically retrieving the data.

Finally the ontology based search systems which provide more precise search for the user query. Ontology query languages like RDQL (Searborne, 2004), SPARQL (prud'hommeaux & Searborne, 2006) are used to query ontology. These systems with its high demand assumed to provide 100% more precise content to the user. The ontology will be created from the heterogeneous environments which contain unstructured text and media contents. The heterogeneous environment will contain all kinds of file formats.

2.1 Ontology

Ontology describes about the concepts, properties describing features and attributes. Ontology has a set of instances of classes which creates the knowledge base. It provides sharing of common knowledge among the user and the systems. Taxonomies which classify the things in hierarchical form are the basic for the creation of ontologies.

Components of ontology include:

- Individuals: are the basic objects.
- Classes: collection of concepts.
- Attributes: Aspects the objects are having.
- Relations: Individuals and classes are related.
- Rules: Logical inferences
- Axioms: overall ontological theory.
- Events: Relational change.

Domain and its properties can be describes by the above components. These ontologies will have structural similarity too. Ontologies can be encoded by ontology languages. DML, SWRL, OWL are the languages for ontology. Creating and manipulating of ontology can be done by the Ontology editors (7).

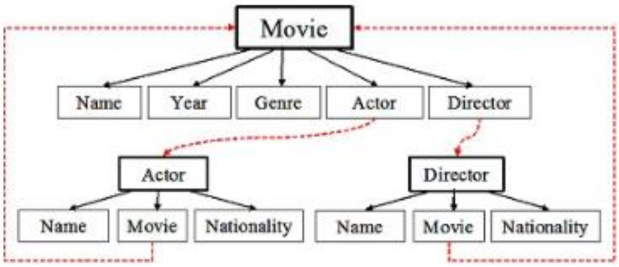

Figure.1 Ontology structure for a movie website

In figure 1, Movie is the class which represents the attributes like name, year, genre, actor, director.

3. METHODOLOGY

The methodology that we have applied here is for the heterogeneous repository of data can be summarized here:

3.1 Critical success factor

Critical success factor (CSF)(1) can be formulated by finding the exact terms that are needed for the retrieval of the data. The solidity of the CSF in the semantic application can be known from the retrieved data’s relevancy. CSF is the core factors that lead to the success of any organization, business... Here the CSF is considered to be the keyword that will support in retrieval of information. It’s the cornerstone in making the query meaningful.

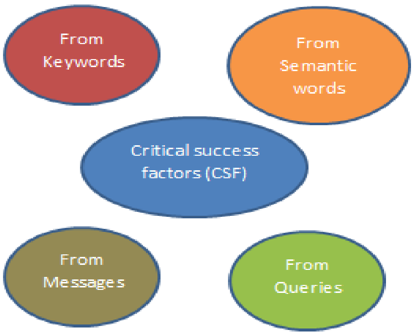

Figure: 2 Identification of CSF

The above fig, describes about the terms from where the CSF will be extracted for the retrieval process.

3.2 Evaluation of clustering:

The application value can be determined by the clustering characteristics .For the grouping of the similar values clustering method is applied to the process. If the records are high we can evaluate statistically. The basic internal criteria for the cluster are to attain high intra-cluster similarity and low inter cluster similarity. In searching criteria, we can evaluate the clusters depending on the time it takes to search the information needed for the user. This can be done from different clustering algorithms. Here we have done with the k-means and spectral clustering algorithms to prove the efficiency through time. It’s an expensive process, if the user data is going to be high.

3.3 Cluster Building:

Exploration of the data can be done by the clusters. Clusters are grouped between the ontology data and the user query. The intra cluster similarity (4) should be high to get a relevant data from the ontology. Hierarchical representation of classes and its instances will be available in ontologies. To group the similar ontological data for a particular user query we need to cluster them. Depending on the dataset size clustering of the data will be done.

3.4 Spectral Clustering:

Spectral clustering is used in many areas like speech processing and data mining. This algorithm performs well than the older methods of algorithm. It uses the standard linear algorithm (6) to solve the cluster data efficiently. We need to construct a pairwise similarity clusters to construct the similarity between the ontology and the query which the user gave. The similarity cluster which has a high similarity values is more advantageous to match the data that are sparsely located. Here the algorithm doesn’t make any assumptions in the form of clusters (6). Spectral clustering algorithm works well for the larger data set by finding the similarity matrix. It’s an effective algorithm to produce good clusters when it applied with care and logic. The results produced from the similarity matrix are more precise.

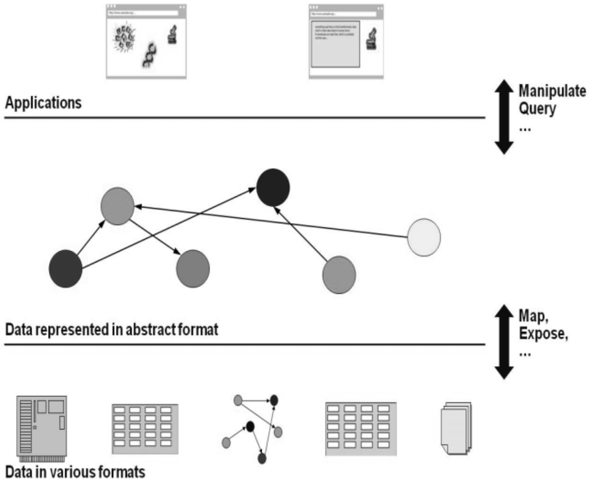

Figure:3 The overall flow of the Semantic information retrieval

Algorithm: Spectral Clustering

```
Input: Data matrix $P\in\mathbb{R}N\times F$ ($N$ =data points,
$F$ = dimensions), $k$ number of clusters
Output: Clusters $Z1,\ldots Zk$ with $Zi=\{i|yi \in hi\}$
Construct pair wise similarity matrix ($i,j$).
Construct degree matrix $D=\mathrm{diag}(d1,\ldots,dN)$
Compute Laplacian $L=D-A$ (unnormalised)
Compute the first $k$ eigen-vectors $u1,\ldots,k$ of $L$
Let $U\in\mathbb{R}N\times k$ contain the vectors $u1,\ldots,k$ as columns
Let $yi\in\mathbb{R}k$ be the vector corresponding to the ith row of $U$
Cluster the points ($yi$) $i=1,\ldots,$ into $k$ clusters $h1,\ldots,hk$ with K-means
```

The figure 3, explains about how data stored in different file formats eg.XML, spreadsheets, flat files, etc. are retrieved. These heterogeneous file formats are extracted and converted to ontologies. These ontologies show the relationship of the data available in the file. The ontologies are than manipulated according to the end user query and provides the result in the applications.

4. MODEL DESIGN

4.1 Framework

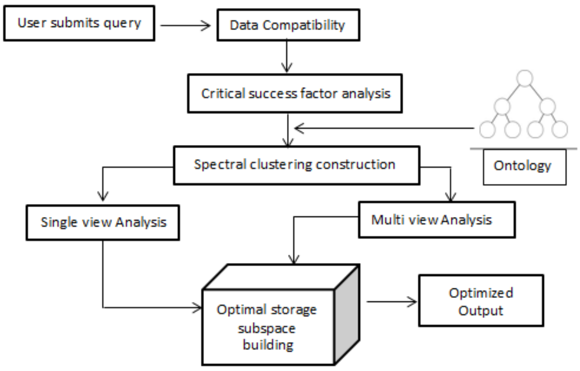

Figure:4 Framework for the semantic retrieval model

The framework provided here in the Fig.4 is the semantic retrieval model and it consists of the following steps:

Step: 1 The user submits the query.

Step: 2 Users query will be entered to the database and then it will be converted into binary attributes to facilitate access to multiple values.

Step: 3 Critical success factor (CSF) are the factors that lead to the success of the semantic applications. The success factors from the user query will be separated. Usually the stop words are removed. Stop words are the commonly used words that are ignored for searching and retrieving data in a search engine. To reduce space and time these stop words are removed while searching and indexing. Adaptive stop word evaluation which will remove the stop words from the list of stop words which are created first.CSF focuses on the important areas which will help to measure success of the process.

Step: 4 Ontology which has to be created from the heterogeneous data formats. Ontology is the softbots which provide repositories of knowledge base about the synonyms, concepts, semantic relationship between concepts and axioms to it.

Step: 5 The CSF will be grouped along with the ontology data with the spectral clustering algorithm. Sparsely located data can be clustered effectively by spectral clustering algorithm. This algorithm is a best method to analyze and cluster data by using the matrices of pairwise distances.

Step: 6 Optimal storage subspace which is a buffered area to store the clustered data. The clustered data is the optimized data that has been retrieved according to the user query.

Step: 7 Finally, the optimized output from the storage will be given to the user.

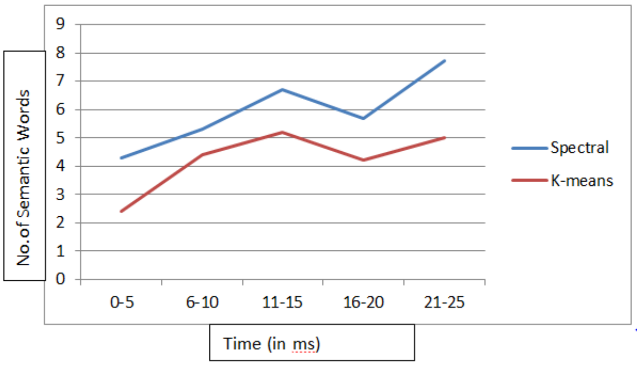

Figure 5: Comparison graph of k-means and spectral algorithm.

The above result shows that the number of related word that are retrieved from both the k-means and spectral clustering algorithms. Spectral algorithm outperforms k-means algorithm by retrieving more relevant data in a limited time.

5. CONCLUSION

The described attempts here in this paper paves way to link the huge amount of unstructured documents in the web to a semantically more relevant data. The central goal of this model is to improve the retrieval performance from the traditional keyword based search.

This novel proposed model integrates the semantic knowledge in the form ontologies. Also it shows that the data retrieved by the spectral algorithm has more relevancy than the traditional algorithms. Semantic retrieval model has also its limitations and challenges with still the lack of semantic knowledge which ended up in inaccurate retrieving of results. And also we need to highlight on the huge size of the web which is endlessly increasing each and every second. As told in this paper, the topic of semantic search is wider and need to be addressed with more future unsolved research lines.

To conclude, in this paper we have given a semantic search model which addresses the heterogeneous environment to certain level. It also overcomes the basic issues with the keyword based retrieval. Future work is open in the field of exploiting more affluent semantic information for providing more coverage on the users information needs and to reduce the cost of retrieving a huge database within the stipulated time.

REFERENCES

[1] Marek Nekvasil a, Vojtěch Svátek b,∗a Adastra.Towards savvy adoption of semantic technology: From published usecases to category-specific adopter readiness models Business Consulting, Karolinská 654/2, 186 00, Praha 8 - Karlín, Czech Republic b University of Economics, Prague, Nám. W. Churchilla 4, 130 67, Praha 3, Czech Republic.

[2] Miriam Fernán dez1, Iván Cantador2, Vanesa López1, David Vallet2, Pablo Castells2, Enrico Motta Semantically enhanced Information Retrieval: an ontology-based approach Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom 2 Departamento de Ingeniería Informática, Universidad Autónoma de Madrid, Madrid, Spain

[3] Guha, R. V., McCool, R., & Miller, E. (2003). Semantic search. In Proceedings of the 12th International World Wide Web Conference (WWW 2003), pp. 700-709. Budapest, Hungary.

[4] Seaborne, A. (2004). RDQL – A Query Language for RDF. W3CMember Submission.

[5] Prud'hommeaux, E., & Seaborne, A. (2006). SPARQL Query Language for RDF. W3C Working Draft. Sabou, M., Gracia, J, Angeletou, S., d'Aquin, M., & Motta, E.,(2007). Evaluating the Semantic Web: A Task-based Approach. In Proceedings of the 6th International Semantic Web Conference (ISWC 2007), pp. 423-437. Busan, South Korea.

[6] Ulrike von Luxburg.A Tutorial on Spectral Clustering

Kaushal Giri, Role of Ontology in Semantic Web DESIDOC Journal of Library & Information Technology, Vol. 31, No. 2, March 2011, pp. 116-120