IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-10 E-ISSN: 2347-2693

Semantic Based Intelligent Information Retrieval through Data mining and Ontology

Muqeem Ahmed

Dept. of CS and IT, MANUU, Hyderabad, India

**Corresponding Author: muqeem.ahmed@gmail.com*

Available online at: www.ijcseonline.org

Received: 21/Sep/2017, Revised: 30/Sep/2017, Accepted: 15/Oct/2017, Published: 30/Oct/2017

***Abstract*—**On the Web and other document repositories the amount of content stored and shared keeps increasing steadily and fast that results in well known difficulties and problems when it comes to finding and properly managing information in massive volumes. In the last decade with the development of search engine technologies Striking progress has been achieved, which collect, store and pre-process information worldwide to return relevant resources instantly in response to users' needs. However, users still miss or need considerable effort sometimes to reach their targets, even if the sought information is present in the search space. Currently consolidated content description and query processing techniques for Information Retrieval are based on keywords, and therefore provide limited capabilities to grasp and exploit the conceptualizations involved in user needs and content meanings are the common caust. This involves limitations such as the inability to describe relations between search terms to solve the limitations of keyword based , the idea of concept based information retrieval ,conceptual search, understood as searching or retrieving by meanings rather than keyword or literal strings, has been the focus of a wide body of research in the information retrieval field. The semantic technologies such as XML and ontology can play important role for the development of semantic based information retrieval. This paper is an attempt to develop semantic based Information Retrieval for the exploitation of domain knowledge to support semantic information retrieval search capabilities in large document repositories more intelligently; it explores the use of semantic technologies such as xml and ontology to support more expressive queries and more accurate results. for this we have collected the documents from the different domains and design the tree structures of the documents in the form of xml and ontology and data mining technique such as clustering and then retrieve the information from this structure based on user interest that provide the concept based.

Keywords*—:* Information retrieval,ontology,datamining,semantic web, K-Mean,search

I. INTRODUCTION

The amount of content stored and shared on the Web and other document repositories is increasing fast and continuously. This enlargement results in well known difficulties and problems, such as finding and properly managing all the existing amount of information. Striking progress has been achieved in the last decade with the development of search engine technologies, which collect, store and pre-process this information to return relevant resources in response to users' needs to reach their targets even if the analyze information is present in the search space. Presently consolidated content description and query processing techniques for Information Retrieval are based on keywords are the common cause for this, and therefore in user needs and contents meanings involvement it provide limited capabilities to grasp and exploit the conceptualizations. This involves limitations such as the inability to describe relations between search terms or the weakness to properly cope with linguistic phenomena such as polismy or synonymy. By new idea of conceptual information retrieval the information retrieval research solve the limitations of keyword based information retrievals, search, understood as searching by meanings rather than literal strings [1]. Some of these methodologies depend on measurable strategies that review the co occurrence of terms, and hence don't make utilization of appropriate semantic innovations. Without considering potential problems such as the polis my phenomenon the relations between terms are extracted from term frequencies. Various other information retrieval use the linguistic approach [2] that is similar to the human mind in these approach the level of conceptualization is often shallow and sparse, especially at the level of relations. In the Information Retrieval since the early eighties the idea of supporting a higher-level conceptual understanding of contents and queries has been present [3], but there are some problems in this conceptual information retrieval [4].Certain recent advances in information retrieval research are then mentioned, including the formulation of new probabilistic retrieval models, and the development of automatic document analysis and Boolean query processing techniques. The Semantic Web vision was brought about with the aim of helping automate tasks which require a

certain level of conceptual understanding of the objects involved, or the task itself, and enabling software programs to automatically find, share and combine information and resources in consistent ways. In the semantic web technologies ontologies are the core of these new technologies [5] that are envisioned as key elements to represent conceptual information that can be understood, used and shared among distributed environments to overcome the limitations of present keyword based information retrieval or search in the Information Retrieval context was soon envisaged and has been explored by different researchers in the form of Semantic Web. For the modelling and codifying information the Semantic Web [6] led to quite mature standards.. Today, Semantic Web technologies such as ontologies, XML and data mining techniques become key technologies for intelligent information processing, that providing a framework for sharing conceptual information about a domain. The Web Ontology Language or OWL [7],onceptNet [8] for the documents searching and retrieving will be be able to convert a user demand into set of discrete concepts. The document similarity is computed by combining semantically enrich terms in the documents and in the queries between the clusters labels and expand concepts of the query terms. Initially this semantic similarity approach is based on WordNet is used [9] that has emerged as the de facto standard for defining Semantic Web ontologies. So there is gap often referred to as the semantic information retrieval research gap between the data, application and data mining algorithm. The semantic web technology can play an intelligent role to bridge the gap to retrieve the relevant information through the data mining algorithms. for systematic incorporation of domain knowledge in an intelligent data mining environment the Semantic Web technologies that formally represent domain information including structured collection of prior information, inference rules, knowledge enriched datasets etc., could thus develop frameworks. The exploitation of Semantic Web technology and WordNet ontology to support semantic Information Retrieval capabilities in Web documents the user model [10] is acquired by analyzing the user behaviour in the system to record user information that is based on user interests [11]. But above discussed research are at initial stage. There is a need of further enhancement in the semantic technology to retrieve intelligent information based on concept rather than keyword in less time to handle the keyword based and user interest based information retrieval. This paper is an attempt to develop a concept based and user interest based information retrieval interface based on clustering, XML and ontology matching. Since the Semantic Web is distributive, there are a lot of resource descriptions where two concepts within different ontologies are equivalent, but they are described in different terms. It matches those elements, and then bring a more accuracy search result based on more expressive queries to achieve the semantic based information retrieval. So the proposed concept based information retrieval framework will be handled more concept based and user interest based information by using semantic technology and data mining techniques.

II. RELATED WORK

The advances in learning building and data mining advance semantic information mining, which conveys rich semantics to all phases of information mining process. Many research endeavors have bore witness to the benefit of fusing area learning into information mining. Formal semantics encoded in the cosmology is all around organized which is simple for the machine to peruse and process in this manner make it a nature approach to utilize ontologies in semantic information mining. Utilizing ontologies, semantic information mining has points of interest to connect semantic holes between the information, applications, information mining calculations, and information mining comes about, give the information mining calculation with priori learning which either manages the mining procedure or lessens the pursuit space, and to give a formal approach to speaking to the information mining stream, from information pre processing to mining comes about Information retrieval called learning disclosure from database, is the procedure of nontrivial extraction of understood, already obscure, and possibly helpful data from information [12]. In the previous couple of decades, propels in Data mining systems prompt numerous momentous transformations in information investigation and huge information. Data mining likewise joins methods from insights, manmade brainpower, machine learning, database framework, and numerous different controls to investigate expansive informational collections. Semantic Data Mining alludes to information mining errands that efficiently join area learning, particularly formal semantics, into the procedure. The viability of space information in information mining has been bore witness to in past research endeavors.

Fayyad et al. [13] guaranteed that space information can assume a vital part in all phases of information mining including, information change, highlight decrease, calculation choice, post-preparing, display translation et cetera. Russell and Norvig [15] trusted that an insightful operator (e.g., an information mining framework) must be able to acquire the foundation information and ought to learn learning all the more successfully with the foundation learning. Past semantic information mining research has confirmed the positive impact of area learning on information mining. For instance, the pre processing can profit by space learning that can enable channel to out the repetitive or conflicting information [14], [16]. Amid the looking and example creating process, area learning can fill in as an arrangement of earlier information of requirements to help decrease seek space and guide the inquiry way [17], [18]. Encourage more, the found examples can be gotten out

[19], [20] or made more obvious by encoding them in the formal structure of information designing [21]. To make utilization of space learning in the information mining process, the initial step must record for speaking to and constructing the learning by models that the PC can additionally access and process. The expansion of information designing (KE) has surprisingly improved the group of area learning with strategies that manufacture and utilize space information formally [22]. Metaphysics is one of fruitful information building progresses, which is the express determination of a conceptualization [23], [24]. Regularly, a cosmology is produced to indicate a specific space (e.g., hereditary qualities). Such metaphysics, regularly known as an area cosmology, formally indicates the ideas and connections in that space. The encoded formal semantics in ontologies is fundamentally utilized for viably sharing and reusing of information and information. Noticeable cases of space ontologies incorporate the Gene Ontology (GO [25]), Unified Medical Language System (UMLS [26]), and more than 300 ontologies in the National Center for Biomedical Ontology (NCBO [27]). Research in the zone of the Semantic Web [28] has prompted very develop principles for displaying and classifying area learning. Today, Semantic Web ontologies turn into a key innovation for canny information preparing, giving a system to sharing applied models about an area. The Web Ontology Language (OWL)

[29], which has developed as the true standard for characterizing Semantic Web ontologies, is generally utilized for this reason. The Semantic Web innovations that formally speak to space learning including organized accumulation of earlier data, surmising rules, information enhanced datasets and so forth., could in this way create structures for orderly joining of area information in an astute information mining condition. In this overview paper, we think about the advances and condition of specialty of semantic information mining. We particularly concentrate on the philosophy based methodologies. The cosmology based methodologies for semantic information mining endeavour to make utilization of formal ontologies in the information mining process. This is for the most part accomplished by utilizing the formal meaning of ideas and connections in ontologies as helper data or limitation conditions to manage the information mining process. For instance, in classification, metaphysics can indicate the consistency connections of the arrangement undertaking. By decision out the conflicting hunt space, the grouping undertaking would bring about a superior exactness [30]. Encourage more; organized association of ontologies can fill in as a decent portrayal for the information mining result. For instance, in data extraction and content mining, the separated data can be displayed through the philosophy itself utilizing a cosmology definition dialect (e.g., OWL) [31]. In the previous decade, to deal with and control the enormous information have brought extraordinary discourse up in the information mining group. With the improvement of information building, particularly Semantic Web methods, mining extensive sum, semantics rich, and heterogeneous information rises as an essential research point in the group. The same number of scientists have brought up, work along semantic information mining is still in its beginning period. Philosophy based semantic information mining is by all accounts one of most encouraging methodologies. The real test is to grow more programmed semantic information mining calculations and frameworks by using the full quality of formal philosophy that has very much characterized portrayal dialect, formal semantics, and thinking apparatuses for rationale induction and consistency checking.

III. METHODOLOGY

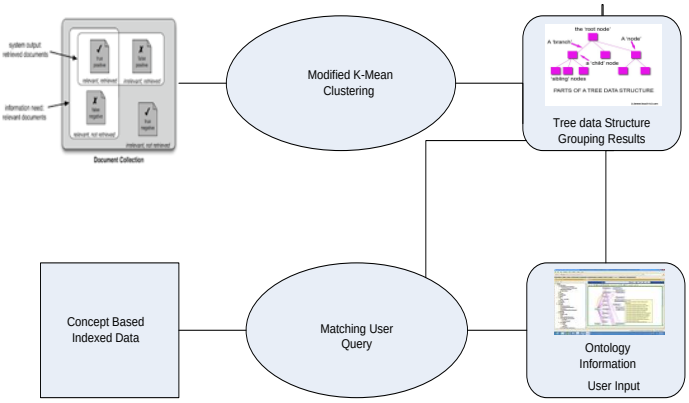

Figure1: Semantic based information Retrieval

Initially, text documents which have been collected from various sources were accumulated in a database. Then, pre-processing was carried out by considering the various stages like tagging by means of Stanford POS tagger tool, stop word removal and stemming, based on Porter Stemmer algorithm and morphological capabilities of WordNet. The above pre processing is common for existing algorithm. The proposed algorithms considered in this study are based on Modified K-Mean clustering. It represented the tree database of collected learning documents. These documents are clustered using modified K-means algorithm which generates K-Number of clusters and the algorithms as originally proposed by the various authors were implemented in the above environment. For uniformity the clustering- based concept semantic similarity alone is implemented with query retrieval and then proceeds to clustering. By varying the number of documents the results of the proposed and existing algorithms are measured. These algorithms are implemented in JDK 1.7 environment using Net Beans IDE

Modified K-means Algorithms for the semantic based text retrieval

For the development intelligent information retrieval the data are grouped into two groups and process k-means algorithm in existing work and modified k-Means algorithm in proposed methodology in the following steps.

K-Means process

eg input: 8,4,6,7,2,1,

Step 1: Initially we select two centroid from whole data because we going to group data into 2 groups cen1=6 cen2=1

Step 2: Find the distance for whole data for these two centroids eg

cen1=6	cen2=1
8	2	7
4	2	3
6	0	5
7	1	6
2	4	1
1	5	0

Step 3: Group the data into two groups with minimum distance

Eg.

group1 8,4,6,7,

group2 2,1

Step 4: update the both centroids

For eg

cen1=4

cen2=2

Step 5: Again find the distance for both centroids

cen1=4	cen2=2
8	4	6
4	0	2
6	2	4
7	3	5
2	2	0
1	3	1

Step 6: Again group the data based on the distance

Group 1: 8,4,6,7

Group 2: 2,1

Step 7: This process repeated until centroid not changed that final group is the result of the k-means algorithm

Final Result

Group 1: 8,4,6,7

Group 2: 2, 1

In this result group one contain 4 items and group2 contains 2 items so both groups are not balance. So to overcome this problem. We used modified k-means algorithms in proposed work.

Modified K-Means algorithm: in modified k-means algorithm for balancing both group we use one technique is Spasis method. Initial process of Spasis method is same as k-means algorithm. After getting final result we change the clustered data on the clustered groups for balancing both groups.

Method 1: neighbour item

Group 1: 8,4,6,7

Group 2: 2, 1

Neighbour item:

Step 1: take first item from cluster 1 eg take item 8

Step 2: find the distance for all items with item 8 from both cluster

8-4=4

8-6=2

8-7=1

2-8=6

1-8=7

Here 7 get the minimum distance item 8 and item 7 are belongs to cluster 1 so here no need to change the cluster Next take the next item here next item is 4 find the distance for 4 with all other items

8-4=4

6-4=2

7-4=3

4-2=2

1=4=3

Here 6 and 2 get the same distance. But both 6 and 2 are belongs to different clusters, so here we need take one decision which cluster have less items cluster 2 contains less items compared to cluster 1So here we moved that 4 from cluster 1 to cluster 2

Group 1: 8, 6, 7

Group 2: 2, 4, 1

This process continued until both cluster reached same number of items

Performance metrics

F-measure and Purity are the performance measures used to evaluate the quality of document clustering. F-measure combines the Precision and Recall from information retrieval process Steinbach et al. (2000). Each cluster is treated as if it were the result of a query, and each class as if it were the desired set of documents, for a query. The recall and precision of that cluster for each given class are calculated. More specifically, F-measure for cluster j and class i is calculated as follows:

Recall (I,j)= nij/ ni(1)

Precision (I,j)= nij /nj(2)

F(i, j)=2*recall(I,j)* Presicion (i j)/Presicion (i j)Recall (i j) (3)

Where nij is the number of members of the class i in cluster j, nj is the number of members of cluster j and ni is the number of members of class i. For each class, only the cluster with highest F-measure is selected. Finally, the overall F-measure of a clustering solution is weighted by the size of each cluster:

$F(s) = 1/ n\sum n^{n}_{j=1}$ nj /max (F (I, j) (4)

The purity measure evaluates the coherence of a cluster, that is, the degree to which a cluster contains documents from a single class Huang (2008). Given a particular cluster Ci of size ni the purity of Ci is formally defined as:

$P\ C_1 = 1 /n \max n^{h}_{i})$ (5)

Where $\max(n_i^h)$ is the number of documents that are from the dominant class in cluster Ci and ni h represents the number

of documents from cluster Ci assigned to class h. The overall purity of a clustering solution is:

$$\text{Purity}(S) = 1/n \sum_{i=1}^{n} \max(n_h^i) \quad (6)$$

Relevant details should be given including experimental design and the technique (s) used along with appropriate statistical methods used clearly along with the year of experimentation (field and laboratory).

IV. RESULTS AND DISCUSSION

The Average transaction size and average maximal potentially frequent item set size are set to maximum. The comparison of cluster results computed using standard K-means algorithm and Modified approach algorithm are different. In the k- mean algorithm as the limitation is the heterogeneity of groups balancing, due to this the time taken to execute the record is more compare to modified k-Means algorithm. So for balancing both the group we used the technique spasis method to overcome this problem. So the time taken to execute the records is less as compare to K-Mean Algorithm. This method follows the initial process same as the k-means algorithm after getting final result we change the clustered data, based on the clustered groups to balance both groups

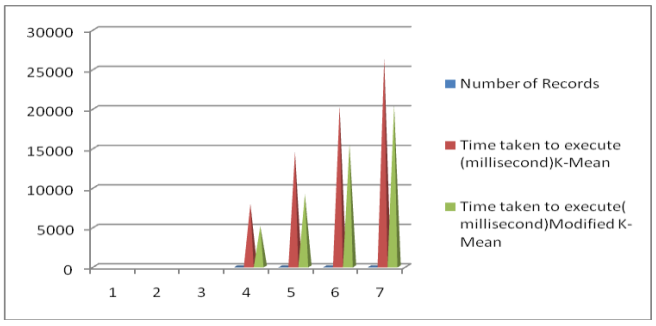

Figure1: Comparison of existing K-means and Modified K-means

The comparison between K-mean and Modified approach K-mean on the basis of large number of records and execution time using this algorithm showed in the above figure. The Modified approach K-mean is better performance comparison to standard K-means algorithm.

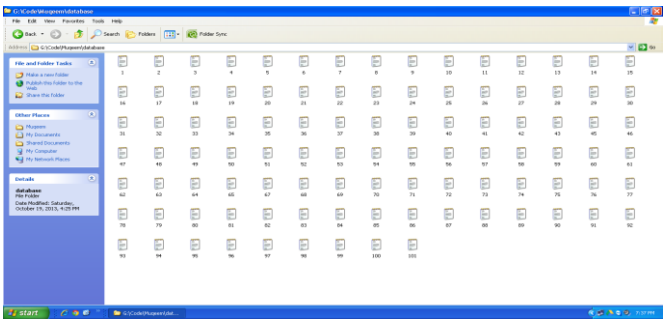

Figure 2: Data set

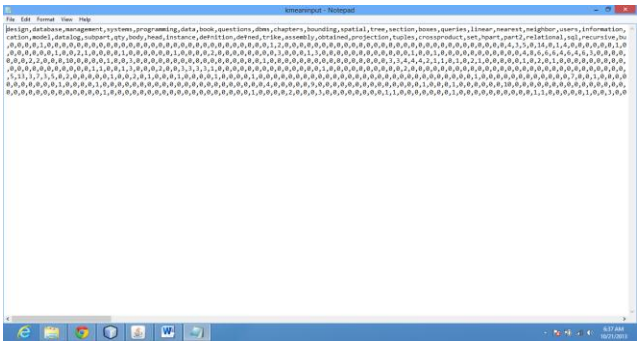

Figure3: Input file

EXPERIMENTAL RESULTS

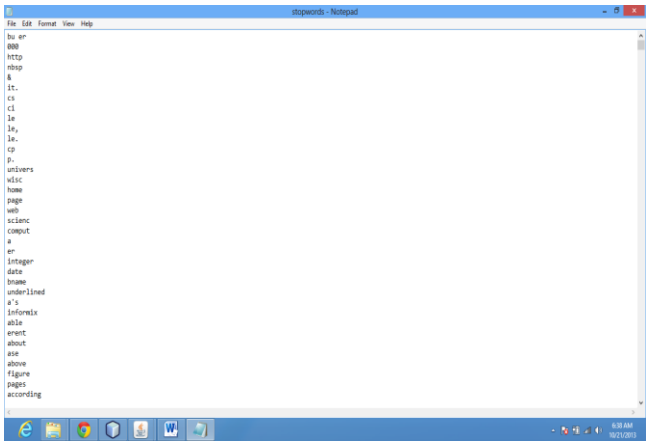

Figure 4: Find best words

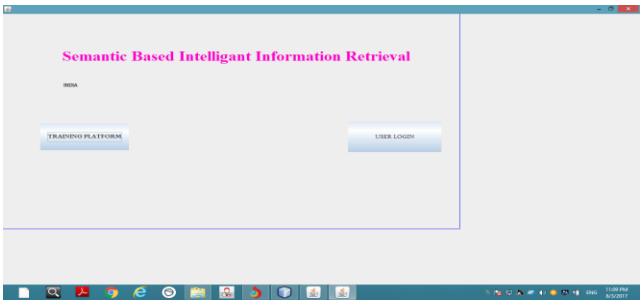

Figure 5: Semantic Based Interface

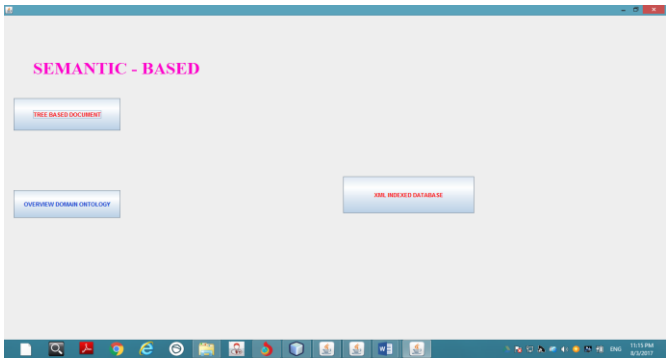

Figure 5.1: Semantic Based Interface

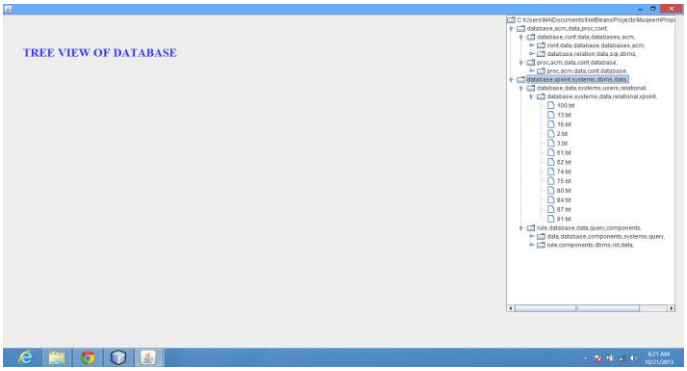

Figure 6: Result Tree data structure of Documents

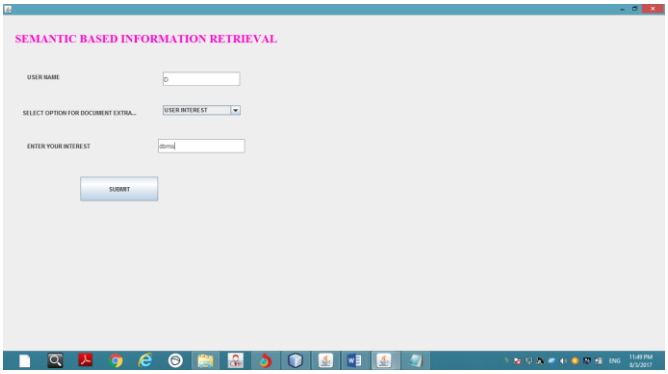

Figure9: Result user view of data

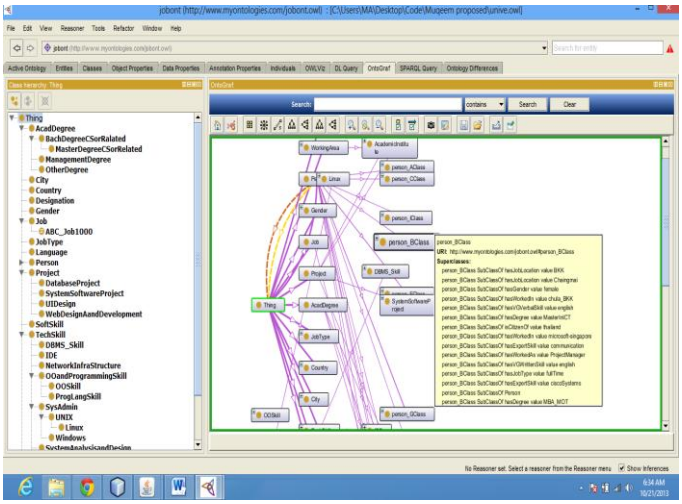

Figure 7: Result Overview of Domain Ontology

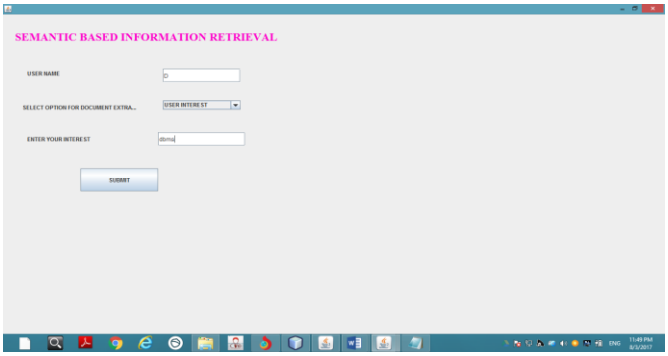

Figure 10: Query Request by users

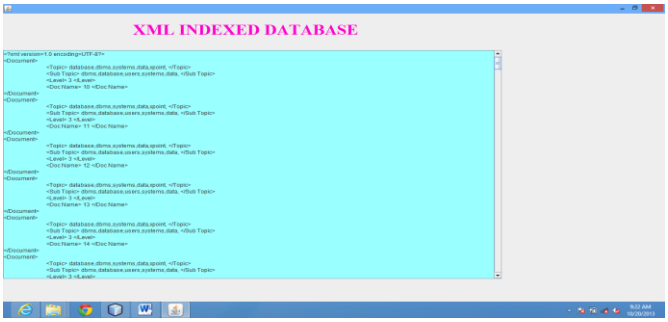

Figure 8: Result xml indexing of data

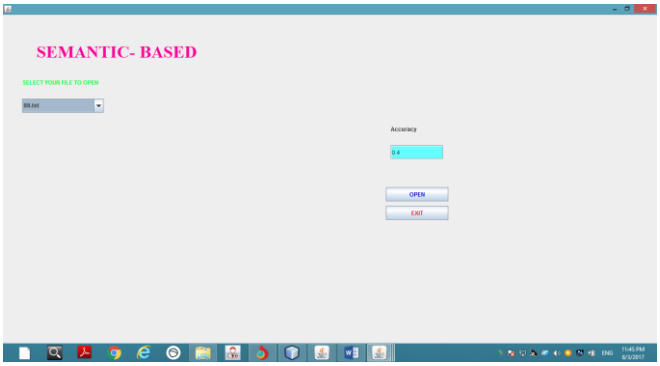

Figure11: Accuracy of Text

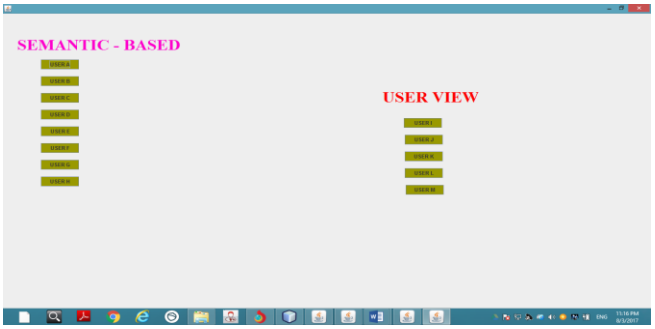

Figure9: Result user view of data

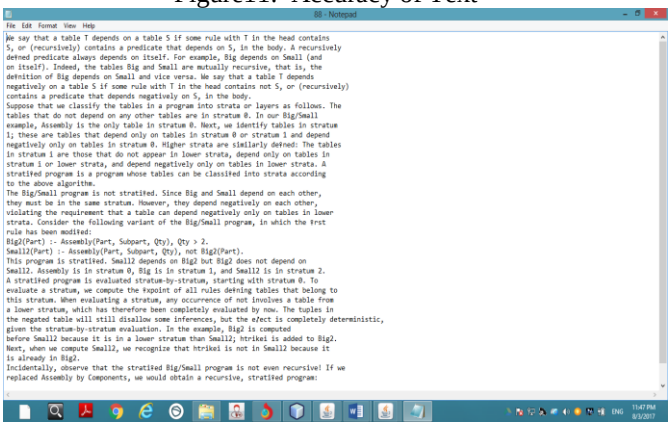

Figure 12: Query Result of information based on concept

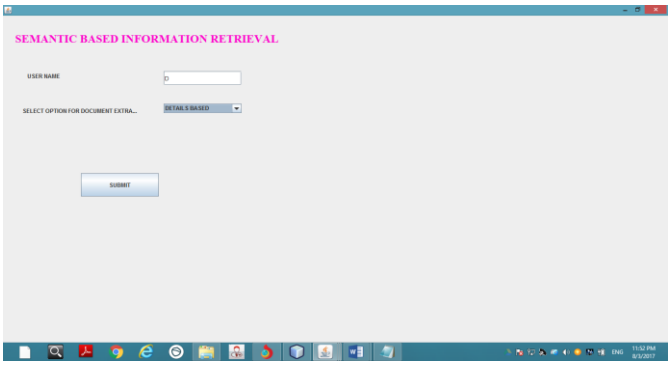

Figure 13: Query Request from user

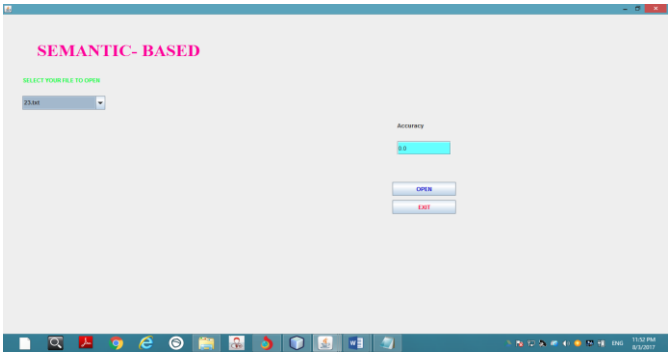

Figure 14: Query Result of information Accuracy

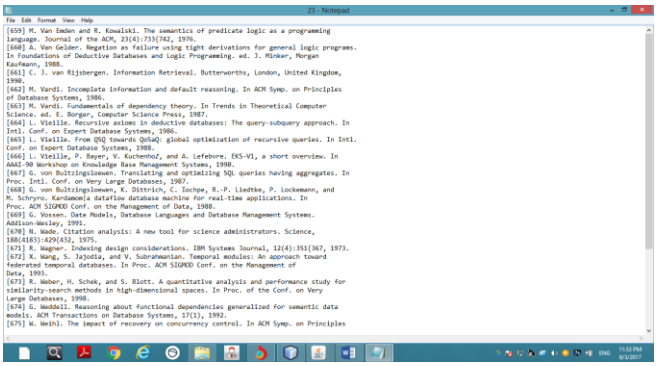
Figure 15: Query Result of information detail based

V. CONCLUSION AND FUTURE SCOPE

Semantic information retrieval is a recent development happened in the field of information retrieval and semantic web due to the growth of information technology. In this platform the challenges involved are to organize the information in a more useful way and also, the retrieval and discovery of useful information from the information space in a more significant way. It has been also pointed out in the above review of literature that present approach has some limitations such as heterogeneity of web based information and the size of data that are responsible for the slow application of semantic information retrieval. Semantic based intelligent information retrieval integrates present information retrieval and semantic web technologies such as the xml and ontology representation of the data that focus information retrieval more intelligently comparison to the previous approach of the information retrieval such as keyword based approach. This approach introduced concept based information by utilizing the semantic technologies. It can also provide better search capability to achieve some qualititative improvement over the keyword based information retrieval by introducing the xml and domain ontology. For the implementation of this we have collected hundreds documents from the different domain and designed the tree data structure in the form of xml and ontology by using java and protégé ontology editor and get the concept based results from the collected documents . As future research line we plan to explore other Semantic Web gateways, to make use of larger amounts of online available semantic metadata. Another important research problem, aside from obtaining high quality knowledge resources, is the annotation of unstructured content. The annotation problem consists in identifying semantic entities within the contents. It is a difficult research problem on its own, which is being widely studied in areas such as Information Retrieval, NLP and Semantic Web

REFERENCES

[1] Susan T. Dumais. George W. Furnas. Thomas K. Landaue" *Indexing by Latent Semantic Analysis*" Bell Communications Research 1990

[2] Gonzalo, Verdejo, Chugur, & Cigarrán "*Sense clusters for information retrieval"* Proceedings of the COLING/ACL'98 Workshop on Usage of WordNet for NLP, Montreal, 1998

[3] Bruce Croft w."*Boolean queries and term dependencies in probabilistic retrieval*" 1986

[4] C. J. van "*Information Retrieval*", ACM sigir Forum, v.17 n.4, 1979

[5] Thomas R.Gruber "A Translation Approach to. Portable Ontology Specifications" . Knowledge Acquisition, 5(2):199-220,1993

[6] T. Berners-Lee, J. Hendler, and O. Lassila. "*The semantic web*". Scientific American, 284(5):28–37, 2001

[7] OWL Web Ontology Language. http://www.w3.org/TR/owl-ref/.

[8] Blanco, E., Cankaya, H. & Moldovan, D. "*Commonsense Knowledge Extraction Using Concepts*" Properties. Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference. 2011

[9] Li, H., Tian, Y., Ye, B. & Cai, Q. " *Comparison of Current Semantic Similarity Methods in WordNet"*. 201O International Conference on Computer Application and System Modeling (ICCASM 2010). 978-1 -4244-7237-6/10, IEEE.

[10] Nidelkou, E., Papastathis, V., Papadogiorgaki, M., Kompatsiaris, I., Bratu, B., Ribiere, M. & Waddington, S. " *User Profile Modeling and Learning"*. In Encyclopedia of Information Science and Technology", Second Edition. DOI: 10.4018/978-1-60566-026- 4.ch627. 3934-3939. IGI Global.

[11] Harb, H., & Fouad, K." *Semantic web based Approach to learn and update Learner Profile in Adaptive E-Learning*." Al-Azhar Engineering Eleventh International Conference, December 23-26 2010

[12] J. Han and M. Kamber. "*Data Mining: Concepts and Techniques*". Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[13] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. *"From data mining to knowledge discovery in databases"*. AI magazine, 17(3):37, 1996.

[14] N. Khasawneh and C.-C. Chan. *"Active user-based and ontology-based web log data preprocessing for web usage mining"*. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pages 325–328, 2006.

[15] S. J. Russell and P. Norvig. *"Artificial Intelligence: A Modern Approach."* Pearson Education, 2 edition, 2003.

[16] D.Perez-Rey, A. Anguita, and J.Crespo.*"Ontodataclean: Ontologybased integration and preprocessing of distributed data. In Biological and Medical Data Analysis*", pages 262–272. Springer, 2006.

[17] N. Balcan, A. Blum, and Y. Mansour.*"Exploiting ontology structures and unlabeled data for learning"*. In Proceedings of the 30th International Conference on Machine Learning, pages 1112–1120, 2013.

[18] A. Bellandi, B. Furletti, V. Grossi, and A. Romei." *Ontology-driven association rule extraction: A case study*". Contexts and Ontologies Representation and Reasoning, page 10, 2007.

[19] C. Marinica and F. Guillet."*Knowledge-based interactive postmining of association rules using ontologies"*. Knowledge and Data Engineering", IEEE Transactions on, 22(6):784–797, 2010.

[20] G. Mansingh, K.-M. Osei-Bryson, and H. Reichgelt."*Using ontologies to facilitate post-processing of association rules by domain experts".* Information Sciences, 181(3):419–434, 2011.

[21] D. C. Wimalasuriya and D. Dou."*Components for information extraction: Ontology-based information extractors and generic platforms"*. In Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM), pages 9–18, 2010.

[22] S. J. Russell and P. Norvig. *"Artificial Intelligence: A Modern Approach"*. Pearson Education, 2 edition, 2003.

[23] T. R. Gruber. "*Toward principles for the design of ontologies used for knowledge sharing"* International journal of human-computer studies, 43(5):907–928, 1995.

[24] R. Studer, V. R. Benjamins, and D. Fensel."*Knowledge engineering: principles and methods". Data & knowledge engineering*, 25(1)':161– 197, 1998.

[25] The gene ontology consortium. *"Creating the gene ontology resource: design and implementation".* Genome Res., 11(8):1425–1433, August 2001.

[26] D. Lindberg, B. Humphries, and A. McCray."*The Unified Medical Language System". Methods of Information in Medicine*, 32(4):281–291, 1993.

[27] *The National Center for Biomedical Ontology. http://www.bioontology.org/.*

[28] T. Berners-Lee, J. Hendler, and O. Lassila. *"The semantic web"*. Scientific American, 284(5):28–37, 2001.

[29] OWL Web Ontology Language. http://www.w3.org/TR/owl-ref/.

[30] N. Balcan, A. Blum, and Y. Mansour." *Exploiting ontology structures and unlabeled data for learning*". In Proceedings of the 30th International Conference on Machine Learning, pages 1112–1120, 2013.

[31] F. Gutierrez, D. Dou, A. Martini, S. Fickas, and H. Zong." *Hybrid ontology-based information extraction for automated text grading"*. In Machine Learning and Applications (ICMLA), 2013 12th International Conference on, volume 1, pages 359–364. IEEE, 2013.

[32] [31] D. C. Wimalasuriya and D. Dou."*Ontology-based information extraction*:" An introduction and a survey of current approaches". Journal of Information Science, 36(3):306–323, 2010.

Authors Profile

Dr. Muqeem Ahmed did Bachelor of Science from Jamia Millia Islaamia India,in 2003 and Master of Computer Application from Jamia Millia Islamia india in year 2006. He did his Ph.D in computer Science from Jamia Millia Islamia Delhi India in 2014. and currently working as Assistant Professor in Department of Computer Sciences, and Information Technology MANUU(Central University) Hyderabad India since 2014. He is a member of IEEE & IEEE computer society since 2014. He has published more than 10 research papers in reputed international journals and conferences including IEEE and it's also available online.He also did a research project funded by UGC. His main research work focuses on Semntic Web and , Big Data Analytics, Data Mining, and Computational Intelligence based education. He has aaround 6 years of teaching experience and 5 years of Research Experience.

