A Survey on-A Semantic Search System for the Vedas

Nikhil Santosh Wani
ME(CSE) student
Department of Computer Science and Engg
GODAVARI COLLEGE OF ENGINEERING, JALGAONJalgaon

Abstract-
The goal of this paper, is to survey on different research papers for, to build the first Semantic Search System for the Vedas, providing normal users and scholars the ability to search the Vedas semantically, analysis all aspects of the text, find hidden patterns and associations using state-of-the-art visualization techniques. And also providing Vedas records in multiple language. That can be use to search and read the Vedas record in multiple language. The major problems raised in existing approaches are fine-grained access, cryptographically access control, measurability in key fine-grained management and effective on-demand user revocation. We would like to provide the secure sharing of Veda record. In this project predominantly considers the multi-owner scenario and divides the Veda record system into multiple security domains that greatly reduces the key management issues. We have to improve the security of Veda Record and set access privileges for every Veda record data. 1. Semantic Search: providing smart semantic search engine for normal users. 2. Visualization: Enhancing the overall visualization of the results and finding new ways to present semantically related data. 3. Question Answering: implementing a question answering system on the top of previous layers. 4. Sentiment Analysis: providing the capability to detect, search by sentiment, and producing the first fully sentiment-labelled Veda corpus.

Keywords-Mining, Clustering.

1. Introduction

The problem lies in the fact that; to implement the goals mentioned earlier, multiple scientific fields and technologies needs to be harnessed and integrated together in one place to serve one purpose. To make a computer respond to user

queries and questions in a smart way and understanding the semantics of both the user input and the target text, the following have to be done: 1. Data should be processed and annotated with as much tags and features as possible, for example the Veda heavily refers to concepts using pronouns, so if there is no corpus to resolve such pronouns the system will miss huge information that is hidden by those pronouns . 2. An Ontology has to be created to describe and link the concepts in the the target domain. This means that ontology extraction from text has to be done in an automated or semi-automated approach which is already an open challenging problem. 3. Custom Question Answering system for the Veda's, has to be implemented based on the ontology. 4. Domain knowledge is needed to understand the text and to facilitate research observations, experiments and evaluation. 5. Much coding, language handling, data model loading and integration, memory/performance optimization and technical experience is needed to implement such system and integrate all modules together. Sound visualization techniques needs to be used to encode and present all semantic information, relations, patterns, insights and answers to the user.

In addition to all the above, since the project is targeting normal users; the online system has to be appealing and usable and self descriptive. Data mining is the process of extracting It makes an important role in data analysis and data mining useful information from the large amount of data and applications. Data divides into similar object groups based on converting it into understandable form for further use. their features, each data group will consist of collection of Clustering is the process of grouping object attributes similar objects in clusters. Clustering is a process of and features such that the data objects in one group are unsupervised learning. Highly superior clusters have high more similar than data objects in another group. But it is intra-class similarity and low inter-class similarity. Several now very challenging due to the sharply increase in the algorithms have been designed to perform clustering, each large volume of data generated by number of one uses different principle. They are divided into applications. Kmeans is a simple and widely used hierarchical, partitioning, density-based, model based algorithm for clustering data. But, the traditional k-algorithms and grid-based. means is computationally expensive; sensitive to

outlier's i.e. unnecessary data and produces unstable result hence Raw Input it becomes inefficient when dealing with very large Data datasets. Solving these Issues is the subject of many recent research works.

2. Literature Survey

Following are general background about all areas researched. Detailed related-work references can be found in the dedicated each topic.

1. NLP & Data Mining- the process or practice of examining large collections of written resources in order to generate new information." The goal of text mining is to discover relevant information in text by transforming the text into data that can be used for further analysis. Text mining accomplishes this through the use of a variety of analysis methodologies; natural language processing (NLP) is one of them.

Although it may sound similar, text mining is very different from the "web search" version of search that most of us are used to, involves serving already known information to a user. Instead, in text mining the main scope is to discover relevant information that is possibly unknown and hidden in the context of other information.NLP-is a component of text mining that performs a special kind of linguistic analysis that essentially helps a machine "read" text. NLP uses a variety of methodologies to decipher the ambiguities in human language, including the following: automatic summarization, part-of-speech tagging, disambiguation, entity-extraction and relations extraction, as well as disambiguation and natural language understanding and recognition. To work, any natural language processing software needs a consistent knowledge base such as a detailed thesaurus, a lexicon of words, a data set for linguistic and grammatical rules, an ontology and up-to-date entities.

2. Semantic Search & Ontology Extraction- Semantic search denotes search with meaning, as distinguished from lexical search where the search engine looks for literal matches of the query words or variants of them, without understanding the overall meaning of the query. Semantic search seeks to improvesearch accuracy by understandingthe searcher intent and thecontexual meaning of terms as they appear in the searchable dataspace, whether on theWeb or within a closed system, to generate more relevant results. Semantic search systems consider various points including context of search, location, intent, variation of words,synonyms generalized and specialized queries,concept matching

and natural language queries to provide relevant search results. Ontology- Ontology is thephilosophical study ofbeing. More broadly, it studies concepts that directly relate to being, in particular becoming, existence,reality, as well as the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology often deals with questions concerning what entities exist or may be said to exist and how such entities may be grouped, related within a hierarchy, and subdivided according to similarities and differences.

Anisha Mariam Thomasa- proposed In his research paper, she described, the text classification method that uses efficient similarity measures to achieve better performance is being proposed in this paper. Semi-supervised clustering is used as a complementary step to text classification and is used to identify the components in text collection. Clustering makes use of labeled texts to capture silhouettes of text clusters and unlabeled texts to adapt its centroids. The category of each text cluster is labeled by the label of texts in it. Thus here the text clustering is used to generate the classification model for the next text classification step. When a new unlabeled text is incoming, measure its similarity with the centroids of the text clusters and give its label with that of the nearest text cluster. The similarity is calculated using different similarity measures. Results and evaluations are summarized and it is found that the system provides better accuracy when a Similarity Measure for Text Processing (SMTP) used for the distance calculation.[1]

Yung-Shen Lin, Jung-Yi Jiang, and Shie-JueLee,- Measuring the similarity between documents is an important operation in the text processing field. In this paper, a new similarity measure is proposed. To compute the similarity between two documents with respect to a feature, the proposed measure takes the following three cases into account: a) The feature appears in both documents, b) the feature appears in only one document, and c) the feature appears in none of the documents. For the first case, the similarity increases as the difference between the two involved feature values decreases. Furthermore, the contribution of the difference is normally scaled. For the second case, a fixed value is contributed to the similarity. For the last case, the feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents.

The effectiveness of our measure is evaluated on several real-world data sets for text classification and clustering problems. The results show that the performance obtained by the proposed measure is better than that achieved by other measures.[2]

Neha Garg, R.K. Gupta,-Due to the current encroachments in technology and also sharp lessening of storage cost, huge extents of documents are being put away in repositories for future references. At the same time, it is time consuming as well as costly to recover the user intrigued documents, out of these gigantic accumulations. Searching of documents can be made more efficient and effective if documents are clustered on the premise of their contents.
This article uncovers a comprehensive discussion on various clustering algorithm used in text mining alongside their merits, demerits and comparisons. Further, author has likewise examined the key challenges of clustering algorithms being used for effective clustering of documents.[3]

It is a franticprocess to identify similar documents or near documents from a huge repository. In this paper, we proposea Greedy algorithmbased on granular computing as a solution to identifythe similar documentsfrom a large collection. The bench mark K-Meansclustering algorithm has been utilized to split the whole dataset into several information granules. The distance between the centroid of each granule and thefeatures or vector generated from the testdocument is measured. The granule with minimum distance is chosen and this granule is again split into another set of granules. This process isrepeated until arriving at a set of documentsoradocumentwhich are/isneareror more similar to the test document. In-order to assess the efficiency of the proposed method, the abstracts of the 100 documents which are related to the research areainmachine learning is taken.[4]

Jaiganesh, S., Jaganathan, P, -Organizing a large volume of documents into categories through clustering facilitates searching and finding the relevant information on the web easier and quicker. Hence we need more efficient clustering algorithms for organizing large volume of documents.Clustering on large text dataset can be effectively done using partitional clustering algorithms. The K-means algorithm is the most suitable partitional clustering approach for handling large volume of data. K-means clustering

algorithm uses a similarity metric that determines the distance from a document to a point that represents a cluster head. This similarity metric plays a vital role in the process of cluster analysis. The usage of suitable similarity metric improves the clustering results. There are varieties of similarity metrics available to find the similarity between any two documents. In this paper, we analyse the performance and effectiveness of these similarity measures in particular to k-means partitional clustering for text document datasets.We use seven text document datasets and five similarity measures namely Euclidean distance, cosine similarity, Jaccard coefficient, Pearson correlation coefficient and Kullback-Leibler Divergence. Based on our experimental study, we conclude that cosine correlation measure is the best suited similarity metric for K-means clustering algorithm.[5]

Sharon X. Lee, Kaleb Leemaqz, proposed, a Finite mixture models have been widely used for the modelling and analysis of data from heterogeneous populations. Maximum likelihood estimation of the parameters is typically carried out via the Expectation-Maximization (EM) algorithm. The complexity of the implementation of the algorithm depends on the parametric distribution that is adopted as the component densities of the mixture model. In the case of the skew normal and skew t-distributions, for example, the E-step would involve complicated expressions that are computationally expensive to evaluate. This can become quite time-consuming for large and/or high-dimensional datasets. In this paper, we develop a multithreaded version of the EM algorithm for the fitting of finite mixture models. Due to the structure of the algorithm for these models, the E- and M-steps can be easily reformulated to be executed in parallel across multiple threads to take advantage of the processing power available in modern-day multicore machines. Our approach is simple and easy to implement, requiring only small changes to standard code. To illustrate the approach, we focus on a fairly general mixture model that includes as special or limiting cases some of the most commonly used mixture models including the normal, t-, skew normal, and skew t-mixture models. The performance gain with our approach is illustrated using two real datasets.[7]

Irwan Bastian, Rozaliyana, MettyMustikasari, Several studies related to the document clustering using K-Means clustering algorithm hadbeen done by some previous studies. Steinbach M, et al (2007)

conducted a study to evaluate two clustering algorithms namely Hierarchical Clustering and K-Means. These results indicatedthat the approach with Hierarchical Clustering better than K-Means. However, a derivative of K-Means such as bisecting K-Means deliveredresults close to the results of Hierarchical Clustering algorithms. This result wasdue to that approach by bisecting K-Means clustering produces significantly fairly consistent.[6]

Huang, proposed similarity measures for text document clustering. She compared and analyzed the effectiveness of these measures in partitional clustering for text document datasets. Her experiments utilizedthe standard K-Means algorithm and she reportedresults on seven text document datasets and five distance/similarity measures that hadbeen most commonly used in text clustering.[8]

Ravindran R.M, Thanamani A.S, proposed K-Means Document Clustering using Vector Space Model. They used Cosine Similarity of Vector Space Model as the centroid for clustering. Using thisapproach,the documents couldbe clustered efficiently even when the dimension was high because it usedvector space representation for documents which wassuitable for high dimensions.[9]

4. Proposed Architecture

TheMulti-Authority Attribute Based Encryption scheme is an advanced attribute based encryption it many attribute authority for handling the different set of users from various domains. In the Veda record system the users will be from different domain like the Teacher that teach the lesson of Veda In this project User Can read the Veda detail information in any language those users want to read. Also we proposed the method that user are search any meaning of word those user want to search simply type the word or sentence. In this project we provide the security at cloud so no one can change the Veda data that are store in the cloud. Also user shares the information to many User. So each user will be having different access control mechanism based on the relation with the User. Thus the MA-ABE scheme will highly reduce the key-management issues and overhead and thus it will provide fine-grained access control to the system. In a multi-authority ABE algorithm consists many attribute authorities and many users. In MA-ABE defines a set of public parameters available to everyone

(cloud server, or by a distributed protocol between the authorities). A user can choose an Attribute Authority (AA), prove that it is entitled to some of the attributes handled by that attribute authority, and request the corresponding distributed key to decrypt Veda data. Architecture diagram for proposed system.

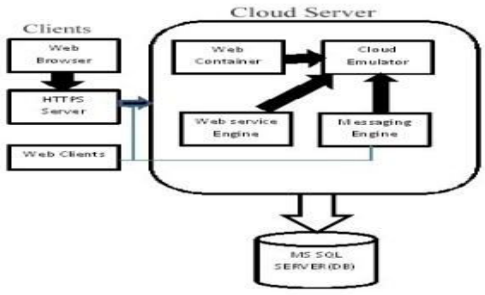

Fig:-Architecture diagram for proposed system.

The multi authority encryption will run the attribute key generation algorithm, and return the result to the user. For each Attribute Authority (AA), a user must have received from each user policy it allows to decrypt a data with set of attributes. The main challenge in MA-ABE is to guarantee that two colluding users cannot each obtain keys from a different authority, and then pool their keys to decrypt a message that they are not entitled to.

5. Conclusion

Ability Semantic Search and Intelligence System for the Vedas, providing normal users and scholars the ability to search the Vedas semantically, analysis all aspects of the text, find hidden patterns and associations using state-of-the-art visualization techniques. Using this anyone can easily read the Vedas in any language also easily search the meaning of any word or sentence quickly so importance of Vedas and meaning of Vedas any one can easily understand also this information user can share to other user securely. Also the Veda Record system provides high level security against third party server. The Veda information is highly secured function for using Multi Authority-Attribute Based Encryption (MA-ABE). It plays a prominent role because these data are unique. So it can't be easily access by third party server. The ABE method addresses the unique challenges in multiple owner scenarios, in that we significantly reduce the Key complexity. So we adopt a novel based MA-ABE to encrypt Veda information in cloud computing. The major issues in existing method are key complexity, security, On Demand revocation etc. The proposed scheme overcome the major issues by using the MA-ABE file encryption and also increases the security for sharing data in cloud system.

6. References

1.Anisha Mariam Thomasa, Resmipriya M "An Efficient Text Classification Scheme Using Clustering", International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)

2.Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, `A Similarity Measure for Text Classification and Clustering, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014.

3. Neha Garg, R.K. Gupta, " Exploration of various Algorithms for Text Mining", (IJEME-2018)

4.A. Jennath, K. Thangvell, "Greedy Algorithm using K-means to identify similar documents based on Granular Computing", (IJTET-2016)

5. Jaiganesh, S., Jaganathan, P. (2015). "An Appropriate Similarity Measure for K-Means Algorithm in Clustering Web Documents", International Journal for Scientific Research & Development, 3(2), 2015.

6. Irwan Bastian, Rozaliyana, MettyMustikasari,"Web Document clustering system using k-means algorithm", (IJARCSE-2016)

7. Sharon X. Lee, Kaleb Leemaqz," A Simple Parallel EM Algorithm for Statistical Learning via Mixture Models", IEEE conference-2016

8. Huang, Anna. Similarity Measures for Text Document Clustering. Hamilton-New Zealand: Proceedings of The New Zealand Computer Science Research Student Conference. 2008.

9. Ravindran R.M, Thanamani A.S., K–Means Document Clustering using Vector Space Model, Bonfring International Journal of data mining , Vol 5, No.2 July 2015.