A Semantic Search System for the Vedas

Prof.Pramod Gosavi
Head of Department
Department of Computer Science and Engg
Jalgaon

Nikhil Santosh Wani
ME(CSE) student
Department of Computer Science and Engg
Jalgaon

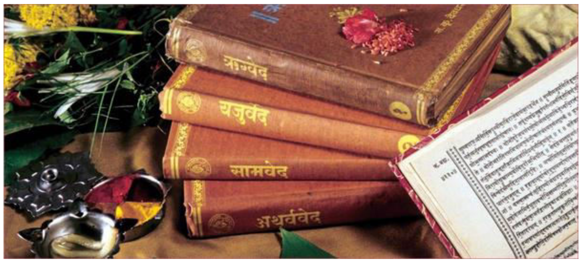
Abstract-
The goal of this paper, to build the first Semantic Search System for the Vedas, providing normal users and scholars the ability to search the Vedas semantically, analysis all aspects of the text, find hidden patterns and associations using state-of-the-art visualization techniques. And also 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

Vedas mean knowledge, a large body of sacred texts originated in ancient India. Vedas are composed in Vedic Sanskrit. Vedas are oldest scriptures of Hinduism. Hindus consider Vedas to be apauruseya which means not of aman and impersonal, authorless. There are four Vedas which are as Rigveda, Yajurveda, Atharvaveda, Samaveda. 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.

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.An Ontology has to be created to describe and link the concepts in 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.Custom Question Answering system for the Vedas, has to be implemented based on the ontology.



Motivation:

The motivation behind the development of this is project is to convert the meaning of Vedas to different languages and people can understand meaning of vedas easily. The language of the Vedas is Sanskrit,so motivation is generated to develop such a system which will minimize the users hard work and time for converting the meaning of the shloka in their regional language.To do so we use text classification algorithm in it.Collection of Data set that is all shloka's of vedas.So to get the proper meaning of each shloka and of each word this system is developed.

Objective:

The main goal of using this system is to convert the meaning of shlokas or hymns in different languages. So that people can understand the vedas effectively and easily. Objective of the system is to translate the given shlokas of the vedas either it is from Rigveda, Yajurveda, Atharvaveda or from Samaveda. People can use system to findtheir solutions. This is the main objective of the semantic search system for vedas.

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 improve search accuracy by understanding the searcher intent and the contexual meaning of terms as they appear in the searchable data space, whether on the Web 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 the philosophical study of being. 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-Jue Lee,- 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 frantic process to identify similar documents or near documents from a huge repository. In this paper, we proposea Greedy algorithm based on granular computing as a solution to identify the similar documents from a large collection. The bench mark K-Means clustering algorithm has been utilized to split the whole dataset into several information granules. The distance between the centroid of each granule and the features or vector generated from the test document 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 documents or a document which are/is nearer or 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, Metty Mustikasari, 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 indicated that the approach with Hierarchical Clustering better than K-Means. However, a derivative of K-Means such as bisecting K-Means delivered results close to the results of Hierarchical Clustering algorithms. This result was due 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 utilized the standard K-Means algorithm and she reported results on seven text document datasets and five distance/similarity measures that had been 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]

3. Proposed Architecture

1.Clustering Algorithms for Text Mining:

With rapid development in technology and in addition sharp lessening in the storage cost, it has turned out to be conceivable to store extensive number of text _les for future references. At the same time, it is time consuming as well as costly to recover the document of interest out of these huge accumulations.

Clustering is a convenient text mining technique that subdivides the documents into desired number of clusters, so that documents in same gathering have higher similarity in contrast with documents having a place with various gathering. The persistence of clustering the documents is to introduce an order by grouping them. At a point, when a gathering is sorted out into clusters, it is easier to recover the required documents.

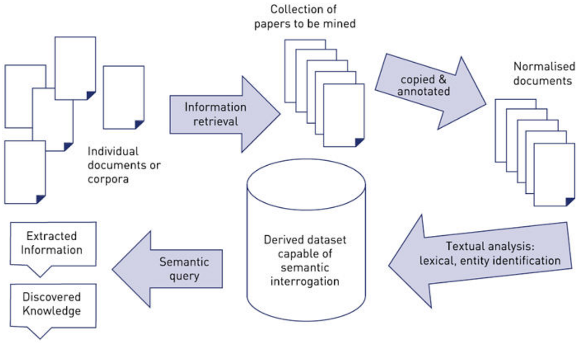

Collection of papers to be mined
Information retrieval
copied & annotated
Normalised documents
Individual documents or corpora
Derived dataset capable of semantic interrogation
Textual analysis: lexical, entity identification
Extracted Information
Semantic query
Discovered Knowledge

Fig:-Architecture diagram for proposed system.

In this algorithm, first to define classes and plot the points on to the coordinates in n-dimensional space.After that using hyper-planes it should be segregate in two classes with maximum margins in it.In non linear fashion also it works properly for that there is an formula to solve that non linear problem. In this way the support vector machine algorithm works in machine learning to solve
the problems.

2.Text Classification Algorithm:

Here dealing with large amount of data today like text, image, and spatial form. So there is great significance of text mining process now a days. There exist many algorithms for text classification but they are having several drawbacks like accuracy, time consumption etc. For overcoming these i am using dimension reduction technique like SMTP, Auto encoder, PCA etc. In this technique creating several clusters and similarity measures are used for calculating similarity of new input document and created clusters. Clustering makes use of labeled texts to capture images of text clusters and unlabeled text to adopt its centroids.

While the similarity is calculated, the clusters that matches the best to the input documents will get that document in it.
Text mining is similar to data mining, here the data is in the form of text. Ituse this when it need to get data from set of text documents. Cluster are groupof similar item sets.Also can make clusters on the basis of distance between nodesor on the basis of similarity measures. This process of making cluster is known as clustering. These things when brought together can make a new system which term as Text Classiffcation Using Clustering. Document is a template.But the thing that leads to the failure of system or result in drawback is calculating inaccurate similarity value, less effciency, Time taken for complete processing, No manual changes in clusters by user. Considering these drawbacks will be designing a system that will overcome these drawbacks. In system will be using more effcient and accurate function known as SMTP. Also it will be
preprocessing the Document which we will take as input. Processing will consist of Extraction of document and Stop word removal.Stop word removal will reduce the time taken for further processing and give
better time complexity. This can be done by Dimension Reduction Technique,Document term matrix, SMTP based similarity measures, Matching SMTP and forming cluster from input query document.

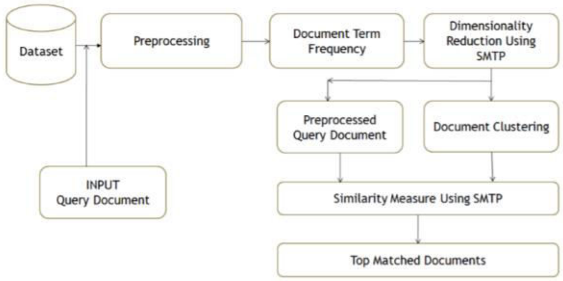


Figure for Text Classification Algorithm

STOP WORD REMOVAL:

Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words are critical to many applications is that, if remove the words that are very commonly used in a given language, can focus on the important words instead. Stop words are language speciffc words which carry no information. The most commonly used stop words in English are
e.g.: is, a, the etc.

3.Expectation Maximization Algorithm:

In this thesis, presenting a new multi resolution algorithm that extends the well-known Expectation Maximization (EM) algorithm for image segmentation.The conventional EM algorithm has prevailed many other segmentation algorithms because of its simplicity and performance. However, it is found to behighly sensitive to noise. To overcome the drawbacks of the EM algorithm we propose a multi resolution algorithm which proved more accurate segmentation than the EM algorithm.
The Expectation-Maximization (EM) algorithm is a way to find maximum-likelihood estimates for model parameters when your data is incomplete, has missing data points, or has unobserved (hidden) latent variables. The EM algorithm can be used to estimate latent variables, like ones that come from mixture distributions (you know they came from a mixture, but not which speciffic distribution).
It is an iterative way to approximate the maximum likelihood function. While maximum likelihood estimation can find the best _t model for a set of data, it doesnt work particularly well for incomplete data sets.The more complex E Malgorithm can find model parameters even if you have

missing data. It works by choosing random values for the missing data points, and using those guesses
to estimate a second set of data. The new values are used to create a better guess for the first set, and the process continues until the algorithm converges on a xed point.Magnetic resonance imaging represents the intensity variation of radio waves generated by biological systems when exposed to radio frequency pulses. A Magnetic resonance image of the human brain is divided into three regions other than the background, white matter, gray matter, and cerebrospinaluid or vasculature. Because most brain structures are anatomically defined by boundaries of these tissues classes, a method to segment tissues into these categories is an important step in quantitative morphology of the brain.

K-Means Clustering Algorithm:

Data mining is the process of extracting useful information from the large amount of data and converting it into understandable form for further use. Clustering is the process of grouping object attributes and features such that the data objects in one group are more similar than data objects in another group. But it is now very challenging due to the sharply increase in the large volume of
data generated by number of applications. K Means is a simple and widely used algorithm for clustering data. But, the traditional K Means is computationally expensive; sensitive to outliers i.e. unnecessary data and produces unstable result hence it becomes inefficient when dealing with very large datasets.Big Data is evolving term that describes any voluminous amount of structured, semi-structured and unstructured data. It is characterized by 5Vs, volume (size of data set), variety (range of data type and source), velocity (speed ofdata in and out), value (how useful the data is), and veracity (quality of data).It creates challenges in their collection, processing, management and analysis.As new data and updates are constantly arriving, there is need of data min into tackle challenges.
Clustering:It makes an important role in data analysis and data mining applications. Data divides into similar object groups based on their features, each data group will consist of collection of similar objects in clusters. Clustering is a process of unsupervised learning. Highly superior clusters have high intra-classsimilarity and low inter-class similarity. Several algorithms have been designed to perform clustering, each one uses diffierent principle.
They are divided into hierarchical, partitioning, density-based, model based
algorithms and grid-based.

Hierarchical Clustering:

A set of nested clusters organized in the form of tree.

Partitioning Clustering:

A division of data objects into subsets (clusters) such that each data object is in exactly one subset
K-means clustering technique is widely used clustering algorithm, which is most popular clustering algorithm that is used in scientific and industrial applications. It is a method of cluster analysis which is used to partition N objects into k clusters in such a way that each object belongs to the cluster with thenearest mean. The Traditional K-Means algorithm is very simple :

1. Select the value of K i.e. Initial centroids.
2. Repeat step 3 and 4 for all data points in dataset.
3. Find the nearest point from that centroids in the Dataset.

4. Form K cluster by assigning each point to its closest centroid.
5. Calculate the new global centroid for each cluster.

K-means is the most commonly used partitioning algorithm in cluster analysis because of its simplicity and performance. But it has some restrictions whendealing with very large datasets because of high computational complexity, sensitive to outliers and its results depends on initial centroids, which are selected randomly. Many solutions have been proposed to improve the performance ofK Means. But no one provide a global solution. Some of proposed algorithms are fast but they fail to maintain the quality of clusters.Some generate clusters of good quality but they are very expensive in term of computational complexity. The outliers are major problem that will effect on quality of clusters. Some algorithm only works on only numerical datasets.

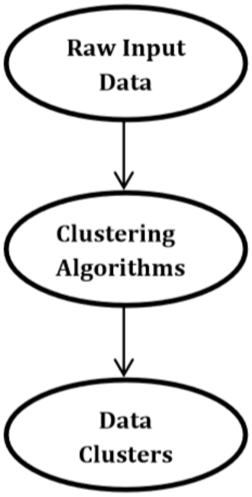


Figure for Stages of K Means Clustering Algorithm

k-means one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters)xed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible faraway from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center.

When no point is pending, the first step is completed and an early group ageis done. At this point we need to re-calculate k new centroids as bary center of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data set points and the nearest newcenter.

A loop has been generated. As a result of this loop that may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function.

4. Conclusion

Algorithm uses the Gaussian filter and the distinct block operation to generate low resolution images from the original image, where two images generated at two successive scales, the parent and the grandparent images. The proposed algorithm has been tested using both synthetic data and manually segmented magnetic resonance images. Moreover, performance analysis between this algorithm and the conventional EM algorithm has been presented.Found that the accuracy of the segmentation done by the proposed algorithm increased signicantly over that of the conventional EM algorithm.

It conclude that there is great importance of text mining process as dealing with large amount of data like text, image and spatial form so calculating similarity measure between the documents. Before the similarity measure, the document should go through various phases of data preprocessing such as extraction, stop word removal, stemming and vector representation. The application is effective for this algorithm. Finally it conclude that through many algorithms have been proposed for clustering but it is still an open problem and looking at the rate at which the web is growing. So SMTP based similarity measure technique for clustering has good effciency for text mining. Most widely used k-means clustering techniques of data mining is analyzed.

This work shows that there are several methods to improve the clustering with different approaches. Various clustering techniques are reviewed which improve the existing algorithm with di_erent perspective. Some limitations of existing algorithm will be eliminated in future work. This technique will be useful in extraction of useful information using cluster from huge database.

5. 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, Metty Mustikasari,"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.