International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 7865-7873
© Research India Publications
http://www.ripublication.com

Text Information Retrieval Using Data Mining Clustering Technique

D.Saravanan
Associate Professor
IBS University, Hyderabad, India.

Abstract

In the aspects of mining, it is used to extract the data's in the efficient manner to reduce the searching time of the user. In data mining cluster play an important role to group the information effectively. The current aspect is to clustering the sentence level text by using the Fuzzy Relational Clustering Algorithm .By means of that, it allows patterns to all clusters .we give a text in a sentence or a sentence that has to be relatively present in documents or a set of documents. An advantage of Fuzzy Relational Clustering Algorithm is that it assigns soft membership values, which can be interpreted as a measure of the degree to which an object belongs to each of the clusters. It is a generic fuzzy clustering algorithm that can be principle be applied to any relational clustering problem, and application to several non-sentence data sets has shown its performance to be comparable to Spectral Clustering. This algorithm that operates on relational input data. Thus the results of applying the algorithm to overlapping clusters of semantically related sentences.it operate the data in the form of similarities. Between the data and the data objects by using the novel fuzzy clustering pattern.

Key terms: Data mining, Clustering, Fuzzy Clustering, Spectral Clustering, Sentence level clustering.

Introduction

Sentence Clustering

Sentence clustering aims at grouping sentences with similar meanings into clusters; commonly, vector similarity measures, such as cosine, are used to define the level of similarity over bag-of-words encoding of the sentences. Then, standard clustering algorithms can be applied to group sentences into clusters. The main challenge for any sentence clustering approach is language variability, where the same meaning can also be listed in different technique.

Sentence level clustering is an application of text classification. The most common objectives in text classification are to classify texts into fairly objective categories such as topics; Clustering has become an increasingly important topic with the explosion of information available via the Internet. Clustering is important it used to group the information to reduce the searching time. Its ability to automatically group similar textual objects together enables one to discover similarity between the given inputs.

Methods used for text clustering include Market Analysis, Financial for costing, grouping data by using logics and simple rule-based systems among the text. In text clustering, it is important to gives the input text properly.

In this paper we focus the text clustering base on two features.

1. Generating relevant term clusters based on lexical semantic relatedness.
2. Projecting the sentence set over these term clusters.

Steps To Obtain Term Cluster:

In order to obtain term clusters, a term connectivity graph is constructed for the given sentence set and is clustered as follows:

1. Create initially an undirected graph with sentence-set terms as nodes and use lexical resources to extract semantically-related terms for each node.
2. Augment the graph nodes with the extracted terms and connect semantically-related nodes with edges. Then, partition the graph into term clusters through a graph clustering algorithm.

Projecting Sentence Ot Term Cluster

To obtain sentence clusters, the given sentence set has to be projected in some manner over the term clusters obtained in Step 1. Our projection procedure resembles unsupervised text categorization (Gliozzo et al., 2005); with categories represented by term clusters that are not predefined but rather emerge from the analyzed data:

1. Represent term clusters and sentences as vectors in term space and calculate the similarity of each sentence with each of the term clusters.
2. Assign each sentence to the best-scoring term cluster. (We focus on hard clustering, but the procedure can be adapted for soft clustering).

Literature Review

Pedrycz, W, Winnipeg, Man, Waletzky, J. et al.[1]proposed a **"Fuzzy clustering with partial supervision "**Presented here is a problem of fuzzy clustering with partial supervision, i.e., unsupervised learning completed in the presence of some labeled patterns. The classification information is incorporated additively as a part of an objective function utilized in the standard FUZZY ISODATA.

Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockettet al. [2]proposed a **"Sentence Similarity Based on Semantic Nets and Corpus Statistics"** This paper focuses directly on computing the similarity between very short texts of sentence length. It presents an algorithm that takes account of semantic information and word order information implied in the sentences. The semantic similarity of two sentences is calculated using information from a structured lexical database and from corpus statistics. The use of a lexical database enables our method to model human common sense knowledge and the incorporation of corpus statistics allows our method to be adaptable to different domains. The proposed method can be used in a variety of applications that involve text knowledge representation and discovery.

Lu, Shin-Yee, Fu, King Sun, et al. [3] proposed an "**A Sentence-to-Sentence Clustering Procedure for Pattern Analysis"** Cluster analysis for patterns represented by sentences is investigated. The similarity between patterns is expressed in terms of the distance between their corresponding sentences. A weighted distance between two strings is defined and its probabilistic interpretation given. The class membership of an input pattern (sentence) is determined according to the nearest neighbor or k-nearest neighbor rule. A clustering procedure on a sentence-to-sentence basis is proposed

Xiaoyan Cai, Wenjie Li, You Ouyang, et al. [4] proposed a **"Simultaneous ranking and clustering of sentences: a reinforcement approach to multi-document summarization "**Multi-document summarization aims to produce a concise summary that contains salient information from a set of source documents. In this field, sentence ranking has hitherto been the issue of most concern. Since documents often cover a number of topic themes with each theme represented by a cluster of highly related sentences, sentence clustering was recently explored in the literature in order to provide more informative summaries. Existing cluster-based ranking approaches applied clustering and ranking in isolation. As a result, the ranking performance will be inevitably influenced by the clustering result. In this paper, we propose a reinforcement approach that tightly integrates ranking and clustering by mutually and simultaneously updating each other so that the performance of both can be improved. Experimental results on the DUC datasets demonstrate its effectiveness and robustness.

Richard Khoury, et al. [5] proposed a "**Sentence Clustering Using Parts-of-Speech** "Clustering algorithms are used in many Natural Language Processing (NLP) tasks. They have proven to be popular and effective tools to use to discover groups of similar linguistic items. In this exploratory paper, we propose a new clustering algorithm to automatically cluster together similar sentences based on the sentences' part-of-speech syntax. The algorithm generates and merges together the clusters using a syntactic similarity metric based on a hierarchical organization of the parts-of-speech. They demonstrate the features of this algorithm by implementing it in a question type classification system, in order to determine the positive or negative impact of different changes to the algorithm

Extracting and Filtering Related Terms

In a number of lexical resources providing pairs of semantically-related terms; within the suggested scheme, any combination of resources may be utilized. Often resources contain terms, which are semantically-related only in certain contexts. E.g., the words visa a passport are semantically-related when talking about tourism, but cannot be considered related in the banking domain, where visa usually occurs in its credit card sense.

In order to discard irrelevant terms, filtering procedures can be employed. E.g., a simple filtering applicable in most cases of sentence clustering in a specific domain would discard candidate related terms, which do not occur sufficiently frequently in a target-domain corpus. In the example above, this procedure would allow avoiding the insertion of passport as related to visa, when considering the banking domain.

Clustering the graph nodes:

Once the term graph is constructed, a graph clustering algorithm is applied resulting in a partition of the graph nodes (terms) into clusters. The choice of a particular algorithm is a parameter of the scheme. Many clustering algorithms consider the graph's edge weights.

To address this trait, different edge weights can be assigned, reflecting the level of confidence that the two terms are indeed validly related and the reliability of the resource, which suggested the corresponding edge (e.g. Word Net synonyms are commonly considered more reliable than statistical thesauri).

Fuzzy Clustering:

Data Clustering in the process of grouping the element in one class or one label. Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" (all-or-nothing) but "fuzzy" in the same sense as fuzzy logic. In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster.

Problem Definition

General

The purpose of the system analysis phase of a project is to analysis the problem statement and finds the additional requirements in the project. In the system analysis phase, the project team defines the problem statement in the project.

Existing System

Existing system uses novel fuzzy clustering algorithm that operates on relational input data; the input data is form of data in the form of a square surrounding substance of data objects. The algorithm uses a graphical method of the input data, that perform in an Expectation-Maximization framework in which the graph centrality of an object in the diagram is interpreted as likelihood. Output of the algorithm applied to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences. The major disadvantage of this process is time and space complexity because of fuzzy clustering mechanism. The existing system uses page ranking mechanism to avoid time complexity but the result is not efficient.

Disadvantages of Existing System

The results often suffered from instability in the optimization algorithms that were used.

A limitation of existing approach is the high dimensionality introduced by representing objects in terms of their similarity with all other object.

Proposed System

The proposed system implements the fuzzy logic with the clustering algorithm like the existing algorithm but instead of using the page ranking mechanism the proposed system uses the N-GRAM preprocessing mechanism which includes stop word removal, stemming etc. to avoid the space and time complexity.

Advantage of Proposed System

- Relational clustering is achieved
- Accurate Search result
- Increase in search speed
- Identifying overlapping clusters.

Exprimental Setup

Pre-Processing

This remove stop words such as prepositions, pronouns, articles, and irrelevant document metadata. Then, adopted a traditional statistical approach for text mining, in which documents are represented in a vector space model. A dimensionality reduction technique used known as Term Variance (TV) that can increase both the effectiveness and efficiency of clustering algorithms. TV selects a number of attributes (in our case 100 words) that have the greatest variances over the documents.

Calculating the number of clusters:

This project implements the approach consists of getting a set of data partitions with different numbers of clusters. A set of partitions result directly from multiple runs of a partitioned algorithm (e.g., K-means) starting from different numbers and initial positions of the cluster prototypes.

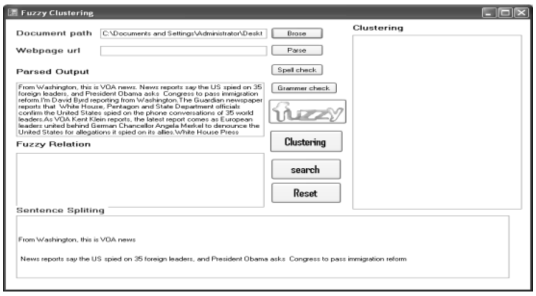
Clustering techniques

In this work partitioned K-means and K-medoids, the hierarchical Single/ Complete/Average Link, and the cluster ensemble based algorithm known as CSPA.For instance, K-medoids is similar to K-means. However, instead of computing centroids, it uses medoids, which are the representative objects of the clusters.

Removing Outliers

Removing Outliers makes recursive use of the silhouette. Fundamentally, if the best partition chosen by the silhouette has singletons (i.e., clusters formed by a single object only), these are removed. Then, the clustering process is repeated over and over again—until a partition without singletons is found. At the end of the process, all singletons are incorporated into the resulting data partition (for evaluation purposes) as single clusters.

Experimental Results



Document path C:\Documents and Settings\Administrator\Deskt

Webpage url

Parsed Output

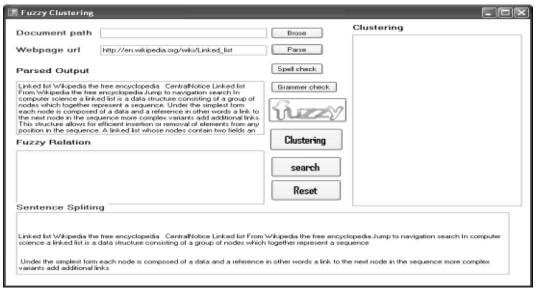
From Washington, this is VOA news. News reports say the US spied on 35 foreign leaders, and President Obama asks Congress to pass immigration reform.I'm David Byrd reporting from Washington.The Guardian newspaper reports that White House, Pentagon and State Department officials confirm the United States spied on the phone conversations of 35 world leaders.As VOA Kent Klein reports, the latest report comes as European leaders united behind German Chancellor Angela Merkel to denounce the United States for allegations it spied on its allies.White House Press

Sentence Spliting

From Washington, this is VOA news

News reports say the US spied on 35 foreign leaders, and President Obama asks Congress to pass immigration reform

Figure 2: Pre-Processing Step



Document path

Webpage url http://en.wikipedia.org/wiki/Linked_list

Parsed Output

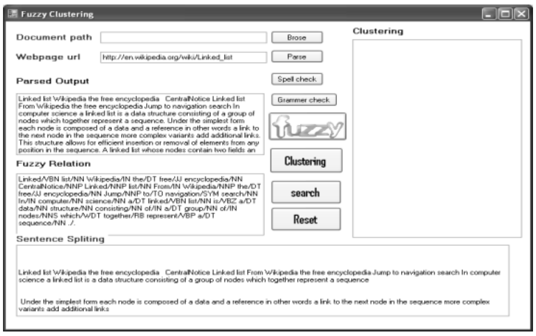
Linked list Wikipedia the free encyclopedia CentralNotice Linked list From Wikipedia the free encyclopedia Jump to navigation search In computer science a linked list is a data structure consisting of a group of nodes which together represent a sequence. Under the simplest form each node is composed of a data and a reference in other words a link to the next node in the sequence more complex variants add additional links. This structure allows for efficient insertion or removal of elements from any position in the sequence. A linked list whose nodes contain two fields an

Sentence Spliting

Linked list Wikipedia the free encyclopedia CentralNotice Linked list From Wikipedia the free encyclopedia Jump to navigation search In computer science a linked list is a data structure consisting of a group of nodes which together represent a sequence

Under the simplest form each node is composed of a data and a reference in other words a link to the next node in the sequence more complex variants add additional links

Figure 3: Calculating the number of clusters



Document path

Webpage url http://en.wikipedia.org/wiki/Linked_list

Parsed Output

Linked list Wikipedia the free encyclopedia CentralNotice Linked list From Wikipedia the free encyclopedia Jump to navigation search In computer science a linked list is a data structure consisting of a group of nodes which together represent a sequence. Under the simplest form each node is composed of a data and a reference in other words a link to the next node in the sequence more complex variants add additional links. This structure allows for efficient insertion or removal of elements from any position in the sequence. A linked list whose nodes contain two fields an

Fuzzy Relation

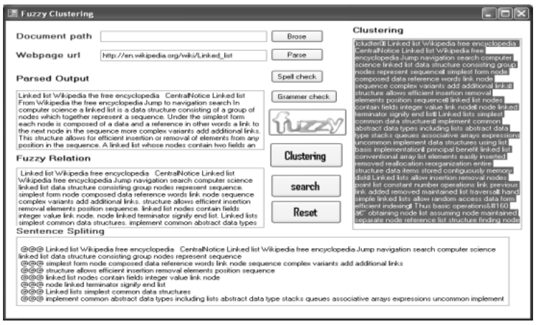
Linked/VBN list/NN Wikipedia/IN the/DT free/JJ encyclopedia/NN CentralNotice/NNP Linked/NNP list/NN From/IN Wikipedia/NNP the/DT free/JJ encyclopedia/NN Jump/NNP to/TO navigation/SYM search/NN In/IN computer/NN science/NN a/DT linked/VBN list/NN is/VBZ a/DT data/NN structure/NN consisting/NN of/IN a/DT group/NN of/IN nodes/NNS which/WDT together/RB represent/VBP a/DT sequence/NN ./.

Sentence Spliting

Linked list Wikipedia the free encyclopedia CentralNotice Linked list From Wikipedia the free encyclopedia Jump to navigation search In computer science a linked list is a data structure consisting of a group of nodes which together represent a sequence

Under the simplest form each node is composed of a data and a reference in other words a link to the next node in the sequence more complex variants add additional links

Figure 4: Clustering techniques



Document path

Webpage url http://en.wikipedia.org/wiki/Linked_list

Parsed Output

Linked list Wikipedia the free encyclopedia CentralNotice Linked list From Wikipedia the free encyclopedia Jump to navigation search In computer science a linked list is a data structure consisting of a group of nodes which together represent a sequence. Under the simplest form each node is composed of a data and a reference in other words a link to the next node in the sequence more complex variants add additional links. This structure allows for efficient insertion or removal of elements from any position in the sequence. A linked list whose nodes contain two fields an

Fuzzy Relation

Linked list Wikipedia free encyclopedia CentralNotice Linked list Wikipedia free encyclopedia Jump navigation search computer science linked list data structure consisting group nodes represent sequence. simplest form node composed data reference words link node sequence complex variants add additional links. structure allows efficient insertion removal elements position sequence. linked list nodes contain fields integer value link node. node linked terminator signify end list. Linked lists simplest common data structures. implement common abstract data types

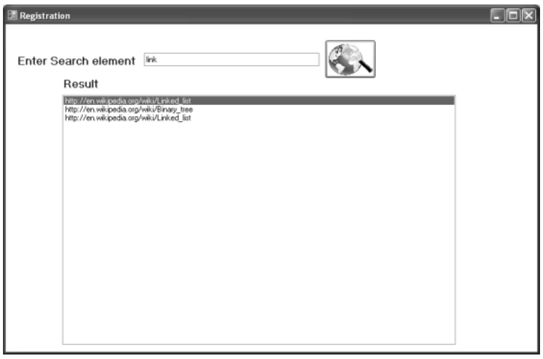
Sentence Spliting

@@@ Linked list Wikipedia free encyclopedia CentralNotice Linked list Wikipedia free encyclopedia Jump navigation search computer science linked list data structure consisting group nodes represent sequence
@@@ simplest form node composed data reference words link node sequence complex variants add additional links
@@@ structure allows efficient insertion removal elements position sequence
@@@ linked list nodes contain fields integer value link node
@@@ node linked terminator signify end list
@@@ Linked lists simplest common data structures
@@@ implement common abstract data types including lists abstract data type stacks queues associative arrays expressions uncommon implement

Clustering

cludterǁ Linked list Wikipedia free encyclopedia CentralNotice Linked list Wikipedia free encyclopedia Jump navigation search computer science linked list data structure consisting group nodes represent sequenceǁ simplest form node composed data reference words link node sequence complex variants add additional linksǁ structure allows efficient insertion removal elements position sequenceǁ linked list nodes contain fields integer value link nodeǁ node linked terminator signify end listǁ Linked lists simplest common data structuresǁ implement common abstract data types including lists abstract data type stacks queues associative arrays expressions uncommon implement data structures using list basis implementationǁ principal benefit linked list conventional array list elements easily inserted removed reallocation reorganization entire structure data items stored contiguously memory diskǁ Linked lists allow insertion removal nodes point list constant number operations link previous link added removed maintained list traversalǁ hand simple linked lists allow random access data form efficient indexingǁ Thus basic operations&#160 â€" obtaining node list assuming node maintained separate node reference list structure finding node

Figure 5: Cluster Formation

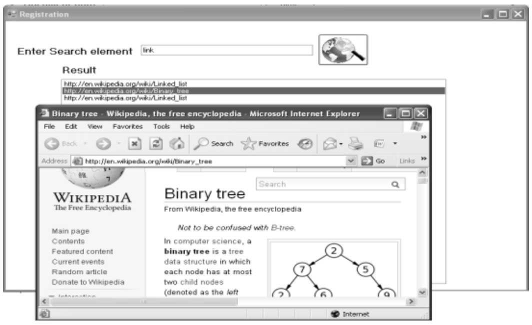


Enter Search element link

Result

http://en.wikipedia.org/wiki/Linked_list
http://en.wikipedia.org/wiki/Binary_tree
http://en.wikipedia.org/wiki/Linked_list

Figure 5: Removing Outliers



Enter Search element link

Result

http://en.wikipedia.org/wiki/Linked_list
http://en.wikipedia.org/wiki/Binary_tree
http://en.wikipedia.org/wiki/Linked_list

http://en.wikipedia.org/wiki/Binary_tree

Binary tree

From Wikipedia, the free encyclopedia

Not to be confused with B-tree.

In computer science, a **binary tree** is a tree data structure in which each node has at most two child nodes (denoted as the *left*

Figure 6: Clstering Search

Conclusion

The results we have presented show that the algorithm is able to achieve superior performance to benchmark Spectral Clustering and k-Medoids algorithms when externally evaluated in hard clustering mode on a challenging data set , and applying the algorithm to a recent news article has demonstrated that the algorithm is capable of identifying overlapping clusters of semantically related sentences. Has shown its performance to be comparable to Spectral Clustering and k-Medoid benchmarks. The algorithm can also be used within more general text mining settings such as user query intended for text mining. Compare with any existing clustering algorithm, performance of the proposed system will ultimately depend on the input given by the user.

Future Enhancement: Any such improvements are orthogonal to the clustering model, and can be easily integrated into it.

Refrences

[1] Pedrycz, W, Winnipeg, Man, Waletzky, J. et al."Fuzzy clustering with partial supervision"Systems,Man and cybernetics, Part B:Cybernetics, IEEE Transactions on volume :27,Issue:5, pp. 787-795..

[2] Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockettet al. **"**Sentence Similarity Based on Semantic Nets and Corpus Statistics"IEEE Transactions on Knowledge and DataEngineering Vol. 18, NO. 8, AUGUST 2006,PP 1138-1150.

[3] Lu, Shin-Yee, Fu, King Sun, et al. proposed an "A Sentence-to-Sentence Clustering Procedure for Pattern Analysis", Pattern Analysis and Machine Intelligence, IEEE Transaction , Vol:28, Issues:2, pp: 302-315.

[4] XiaoyanCai, Wenjie Li, You Ouyang, et al. "Simultaneous ranking and clustering of sentences: a reinforcement approach to multi-document summarization",in the Proc of COLING'10 international conference on computational Linguistics ,pp134-142.

[5] Richard Khoury, et al."Sentence Clustering Using Parts-of-Speech", I.J. Information Engineering and Electronic Business, 2012, 1, pp: 1-9.

[6] D.Saravanan, Dr.S.Srinivasan, "Matrix Based Indexing Technique for Video Data ", International journal of Computer Science". 9(5):534-542,2013, ISSN 1549-3636

[7] V. Hatzivassiloglou, J.L. Klavans, M.L. Holcombe, R. Barzilay, M.Kan, and K.R. McKeown, "SIMFINDER: A Flexible Clustering Tool for Summarization," Proc. NAACL Workshop Automatic Summarization, pp. 41-49, 2001.

[8] H. Zha, "Generic Summarization and Key phrase Extraction Using Mutual Reinforcement Principle and Sentence clustering," Proc.25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 113-120, 2002.

[9] D.R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-Based Summarization of Multiple Documents," Information Processing and Management: An Int'l J., vol. 40, pp. 919-938, 2004.

[10] R.M. Aliguyev, "A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization, "Expert Systems with Applications2009.

[11] R. Kosala and H. Blockeel, "Web Mining Research: A Survey,"ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15, 2000.

[12] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, 1989.

[13] J.B MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” Proc. Fifth Berkeley Symp. Math. Statistics and Probability, pp. 281-297, 1967.

[14] G. Ball and D. Hall, “A Clustering Technique for Summarizing Multivariate Data,” Behavioral Science, vol. 12, pp. 153-155, 1967.

[15] D.Saravanan, Dr.S.Srinivasan, “ A proposed New Algorithm for Hierarchical Clustering suitable for Video Data mining.”, International journal of Data Mining and Knowledge Engineering”, Volume 3, Number 9, july 2011.pp 569-572

[16] J.C. Dunn, “A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters,” J. Cybernetics, vol. 3, no. 3, pp. 32-57, 1973.

[17] R. Mihalcea, C. Corley, and C. Strapparava, “Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity,” Proc.21st Nat’l Conf. Artificial Intelligence, pp. 775-780, 2006.

[18] D. Wang, T. Li, S. Zhu, and C. Ding, “Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization,” Proc. 31st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 307-314, 2008.

[19] D.Saravanan, Dr.S.Srinivasan, “Data Mining Framework for Video Data”, In the Proc.of International Conference on Recent Advances in Space Technology Services & Climate Change (RSTS&CC-2010), held at Sathyabama University, Chennai, November 13-15, 2010.Pages 196-198.

[20] A.Budanitsky and G. Hirst, “Evaluating Word Net-Based Measures of Lexical Semantic Relatedness,” Computational Linguistics, vol. 32, no. 1, pp. 13-47, 2006.

[21] A.Y. Ng, M.I. Jordan, and Y. Weiss, “On Spectral Clustering: Analysis and an Algorithm,” Proc. Advances in Neural Information Processing Systems, pp. 849-856, 2001.

[22] D. Saravanan, Dr.S.Srinivasan, “Video Image Retrieval Using Data Mining Techniques “Journal of Computer Applications, Volume V, Issue No.1. Jan-Mar 2012. Pages39-42. ISSN: 0974-1925