Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Semantic Clustering of Genomic Documents Using Go Terms as Feature Set

V.Bhuvaneswari[1] and Dr. B.L.Shivakumar[2]

[1] Bharathiar University

Received: 13 December 2011 Accepted: 31 December 2011 Published: 15 January 2012

Abstract

The biological databases generate huge volume of genomics and proteomics data. The sequence information is used by researches to find similarity of genes, proteins and to find other related information. The genomic sequence database consists of large number of attributes as annotations, represented for defining the sequences in Xml format. It is necessary to have proper mechanism to group the documents for information retrieval. Data mining techniques like clustering and classification methods can be used to group the documents. The objective of the paper is to analyze the set of keywords which can be represented as features for grouping the documents semantically. This paper focuses on clustering genomic documents based on both structural and content similarity .The structural similarity is found using structural path between the documents. The semantic similarity is found for the structurally similar documents. We have proposed a methodology to cluster the genomic documents using sequence attributes without using the sequence data. The sequence attributes for genomic documents are analyzed using Filter based feature selection methods to find the relevant feature set for grouping the similar documents. Based on the attribute ranking we have clustered the similar documents using All Keyword approach (KBA) and GO Terms based approach (GOTA). The experimental results of the clusters are validated for two approaches by inferring biological meaning using Gene Ontology. From the results it was inferred that all keywords based approach grouped documents based on the semantic meaning of Gene Ontology terms. The GO terms based approach grouped larger number of documents without considering any other keywords, which is semantically relevant which results in reducing the complexity of the attributes considered. We claim that using GO terms can alone be used as features set to group genomic documents with high similarity.

Index terms— Semantic Clustering, Go Terms, Attributes, Feature Set, Xml.

1 Semantic Clustering of Genomic Documents

Using Go Terms as Feature Set Dr.B.L.Shivakumar ? & V.Bhuvaneswari ? Abstract -The biological databases generate huge volume of genomics and proteomics data. The sequence information is used by researches to find similarity of genes, proteins and to find other related information. The genomic sequence database consists of large number of attributes as annotations, represented for defining the sequences in Xml format. It is necessary to have proper mechanism to group the documents for information retrieval. Data mining techniques like clustering and classification methods can be used to group the documents. The objective of the paper is to analyze the set of keywords which can be represented as features for grouping the documents semantically. This paper focuses on clustering genomic documents based on both structural and content similarity .The structural similarity is

found using structural path between the documents. The semantic similarity is found for the structurally similar documents. We have proposed a methodology to cluster the genomic documents using sequence attributes without using the sequence data. The sequence attributes for genomic documents are analyzed using Filter based feature selection methods to find the relevant feature set for grouping the similar documents. Based on the attribute ranking we have clustered the similar documents using All Keyword approach (KBA) and GO Terms based approach (GOTA). The experimental results of the clusters are validated for two approaches by inferring biological meaning using Gene Ontology. From the results it was inferred that all keywords based approach grouped documents based on the semantic meaning of Gene Ontology terms. The GO terms based approach grouped larger number of documents without considering any other keywords, which is semantically relevant which results in reducing the complexity of the attributes considered. We claim that using GO terms can alone be used as features set to group genomic documents with high similarity.

Keywords : Semantic Clustering, Go Terms, Attributes, Feature Set, Xml.

iological data sources are characterized by a very high degree of heterogeneity in terms of the type of data model used, the schema design within a given data model, as well as incompatible formats and nomenclature of values. The biological databases generate huge volumes of genomics and proteomics data after the draft of human genome sequences in 2001. The researchers use the existing sequence information to find similar patterns of genes, proteins and derive other sequence information. Each data source has custom text formats, and these formats change occasionally. Furthermore, an entire data source may be retired or completely restructured using a new schema. Some data sources are inconsistent at the semantic level, and frequently, there is inadequate use of controlled vocabularies and common data elements to specify the metadata.

The National Center for Biotechnology Information (NCBI) is one major resource that maintains public biomedical annotation databases, which are represented in different useful formats that includes XML format. The XML format of databases is very useful, because it is one of the powerful languages for representing the biological data in semi structured form and also the extraction of biological entities from XML format of databases are very easy at any extent. The content similarity measure needs distances that estimate similarity in terms of the textual content inside elements, while the structure dimension needs distances that estimate similarity in terms of the structural relationships of the elements [9].

The Genomic sequence data are stored in public databases like NCBI, Uniport in various formats. The genomic sequence data consist of large number of attributes for describing the sequences. Finding the important attributes for comparing the genomic sequence data based on annotation, becomes the challenging task. Feature selection methods can be used to analyze and study the best features used for representing sequence information for association and clustering of documents.. The complexity of clustering the documents based on the description without considering the sequence data depends on the features selected for clustering. We have analyzed and ranked the features using Filter Based Approach by using CHIR and x 2 statistics.

The Gene Ontology (GO) is one of the most important ontologies in the bioinformatics community and is developed by GO consortium. It is specifically intended in annotating gene products with consistent, controlled and structured vocabulary. The semantic similarity between the documents is determined based on its contents. Many approaches has been used to cluster biological documents based on contents. We have proposed an idea using Gene Ontology terms as a filter to group documents to get meaningful clusters and C compared the same by considering other attributes as keywords leaving GO terms using. In this paper the grouping of biological documents in Xml is done based on structural similarity followed by semantic similarity.

The paper is organized as follows: Section 2 provides the literature review of the study for clustering XML documents and Filter based Feature Selection methods.

Section 3 explains the proposed methodology in detail. Section 4 discussed the experimental results of the proposed work followed by conclusion in Section 5.

The background study related to the work is discussed in detail in the following section. The various approaches to find the similarity between the documents are syntactic similarity and semantic similarity. The related work based on structural and semantic similarity to cluster the documents is as follows. The structural similarity between xml documents is found using graph edit distance measure by Nierman and Jagadish. Edit distance is operations performed on a graph to transform form one form to other [12]. Raffaele has proposed an XML based approach for automatic musicological analysis [14]. Joachim and Paul have presented the use of XQuery with illustrations fop retrieving musical features in music Xml [8].

Tagarelli A and Grew has addressed the problem of clustering xml data based on structure and contents [15]. Ma & Chbili have studied the method for using same schema for finding similarity of XML data based on structure and content [10]. Thedoore and Cheng have proposed a method for clustering XML documents based on structure using tree representation [16] . Docuet A has proposed an approach for clustering homogenous xml documents based on Kmeans algorithm [6]. Panagiotis and Christos has proposed a clustering algorithm for Heterogeneous and homogenous XML using Edge summaries **??**130]. Nayak R has discussed clustering of heterogeneous Xml documents [11].

Bertino has given an matching algorithm for measuring structural similarities between Xml documents and DTD applications [3] . Yu-Chih and Jia has proposed an approach for extraction and clustering structural features for Music XML [19]. Wang [18] proposed a hierarchical algorithm for structural similarity which reduces the join cost for querying XML documents which is stored in relational tables.. The contents in the biological databases

are represented as xml tags. Inferring information from the xml tags which have biological semantic meaning is very complex. The bioinformatics community used various ontologies to infer meaningful biological similarities across documents. The various work proposed by researchers using Go ontology for clustering are as follows: Meeta Mistry and Paul Pavlidas has proposed various content similarity measure using GO and also represented GO as flat matrix representation [9]. Catia Pesquita [5] has evaluated GO based semantic similarity measure using the relationship with sequence similarity as a means to measure based on the presence and absence of these annotations. Brendan and Sheehan [2] has proposed and idea to measure the semantic similarity based on set based and vector based approaches using GO based on conceptual level and structure level. Julie Chabalier ad Jean Mosser [7] has used vector space model for computing semantic similarity between genes using a traversal approach. Andreas Schlicker [1], Francisco has presented a new method for comparing set of GO terms and assessing the functional similarity of gene products. Gene products are said to be functionally similar if they have comparable molecular functions and are involved in a similar biological process.

Feature selection methods have been successfully applied to text categorization but seldom applied to text clustering due to the unavailability of class label information. Bassam Al-Salemi Used Feature Selection techniques such as Mutual Information (MI), Chi-Square Statistic (CHI), Information Gain (IG), GSS Coefficient (GSS) and Odds Ratio (OR) to reduce the dimensionality of feature space by eliminating the features that are considered irrelevant for the category [4].

The proposed methodology shows in Figure **??** 1 consists of two phases for clustering genomic sequence documents using the sequence descriptions. The first phase is the structural similarity phase where the original documents are analyzed based on structure. The filtered structurally similar documents are passed for measuring the content similarity. The second phase the features of the sequence documents are analyzed based on supervised statistical techniques and semantic grouping of documents are done. Two approaches are proposed to group similar documents based on the features. The first approach All Keyword based approach the clusters are analyzed using all the keywords. The second approach GO Terms based clusters analyze the similarity among documents using the GO as keywords. The structural similarity of XML documents is based on the path of the elements given in the document.

The structure of XML document is represented as a tree structure in which it is broken down into collection of distinct paths. The structural similarity is measured using the distinct paths. The sequence database maintains the sequence information as tags in Xml documents. The genomic data in XML format has more than 3500 tags to represent the functional descriptions about the sequences like accession no, taxonomy, organism, lineage, sequence title, sequence descriptions, alternate name, gene name, author details, and identifiers related to other databases like GO, KEGG, PUBMED. To measure the structural similarity between the documents the structural matrix is constructed, in which each document is checked for the below said tags, where there is possibility of more than one occurrence of a particular tag. The total count of occurrence of each tag is entered into the matrix, in absence of a tag value zero is entered into the corresponding place. The content similarities of documents are analyzed only for the documents that are structurally similar. The proposed work the dataset contains sequence attributes for both E.Coli and Human organism. b) Phase Ii Semantic Similarity i.

2 Dataset

In our experiment the public database downloaded from NCBI for E.Coli, human sequencec in xml format is used. The NCBI dataset is the integrated, text-based search and retrieval system used at the major databases, including PubMed, Nucleotide and Protein Sequences, ProteinStructures, Complete Genomes, Taxanomy, and others. The GO Ontology recent download 2010 was used to verify the clusters generated based on the functionality of genes described in the second approach. We have extracted 150 documents from the databases for the organisms and stored in db2 for further extraction and querying. The work is implemented using two softwares. The Xml preprocessing and extraction is carried out using DB2 an IBM Product using XQuery language and we have linked with.NET Framework using COM.

ii.

3 Feature Set Identification

Filter Based approaches supervised methods like X 2 statistics, CHIR statistics are used for analyzing the features of the xml document for the proposed work . The X 2 Statistics is used to measure the independence between the keyword and the category [4]. This can be done by comparing observed frequency in the 2-way contingency table with the expected frequency when they are assumed to be independent. CHIR is a supervised learning algorithm based on X 2 statistics, which determines the dependency between a keyword and a category and also the type of dependency **??**17]. Type of dependency indicates whether the feature is a positive or negative dependency for the category. The Features are Ranked based on X 2 max, X 2 avg and rx 2 statistics. The highly ranked Features are used for analyzing the term relevance. The Feature sets are identified based on ranking

The documents are initially clustered for analyzing the features using hierarchical clustering algorithm for assigning class labels. The proposed work we have considered 150 documents with 358 extracted keywords. On clustering the 150 documents 30 clusters are generated. From the generated cluster it is found that single document is found in many clusters and maximum documents are found in 9 clusters. So we have taken the

cluster which contains highest number of documents to analyze the feature attributes and find the term relevance using filter based approach. Among 358 keywords retrieved we have identified three feature set based on the ranking with 156, 77 and 58 keywords respectively.. The feature set identified are considered for grouping the documents based on its contents.

iii.

4 Semantic Similarity

The content similarity is the main task involved in document clustering, in which the important terms from the documents that differentiate the documents are identified. The term matrix (vector space model) is constructed for the documents which are structurally similar. Consider there are n number of documents in a data set D, that are denoted by d 1 , d 2 , d 3 , ?., d n and the distinct terms from the above document are denoted by t 1 , t 2 , t 3 , ?.., t m . Then the term matrix of size n×m, where n is the number of documents in dataset and m is the number of distinct terms appeared in the data set D, is constructed.. Two different clustering approaches namely All keywords based approach and GO Terms based approach are proposed for clustering similar documents based on the sequence annotations. All keyword based C approach the feature set extracted from the documents are represented as keywords and the term matrix is generated. The documents are clustered using the existing similarity measures like Euclidean, jaccard and cosine . In GO Term based cluster approach the GO terms alone from the feature set are extracted and the mapping is done to find the corresponding genes for the GO terms and viceversa using GO ontology. The term matrix is constructed for the genes and GO terms and the documents are clustered.

5 c) All Keywords Based Approach -Kba

The feature set with 156 keywords and 77 keywords is used as dataset and the feature matrix is constructed. Some biological keywords like Alternate name, Go terms, Gene name, Sp_block keywords and Ecnumber are ranked high in the feature set identification , which has a positive dependency for the clusters generated. The selected feature set attributes were analyzed with respect to the document by varying the no of attributes and clusters were generated. In order to get high degree of cohesion in documents in each clusters we used kernel approach [10], in which the documents in each clusters have high degree of similarity. Kernel is the count of the individual unique keywords from the term matrix greater than a particular threshold. The kernels were created for values starting from 30 and varying it up to 55.The clusters were generated by varying the kernel to find the similarity among attributes. The clusters generated for the kernel values are shown in Figure **??**.

6 d) Go Terms Based Approach -Gota

The proposed idea of our work choosing GO terms as keywords for clustering documents is based on the idea that Documents are said to be semantically based on the gene products. Gene products are said to be functionally similar if they have comparable molecular functions and are involved in similar biological process.

GO annotation capture the available information of genes and used as a basis for defining a measure of functional similarity between genes which is used in our second approach to group documents based on semantic similarity. Each gene is related with more than one GO terms.

A Vector Space Model(VSM) is used to compute similarity between pair of gene products .VSM are essentially used in information retrieval for computing the similarity between documents described as vectors of Keywords **??**2007]. We have used the same model for our second approach to find associative relations between the terms in the GO. To compute the similarity between documents the gene products are described as vectors of GO terms. A gene product is represented by a specific vector g as follows: g=(t 1 ,t 2 ?.t n) where t i is the numeric value that the term takes on for the gene product and n is the number of go terms associated with the gene products. A value t i = 0 means when there exists no association between GO terms and genes. The existing similarity measures are used to cluster the documents.

All the go terms and gene names are extracted from the feature set and a mapping is done with existing go terms and genes using the bioinformatics famous Gene Ontology recently downloaded. The term matrix is constructed representing genes as rows and GO terms as columns for the clustering phase .In the proposed approach we have included only the Go terms as features for clustering leaving other attributes from the documents. We have compared the two approaches and results are discussed in section 4.

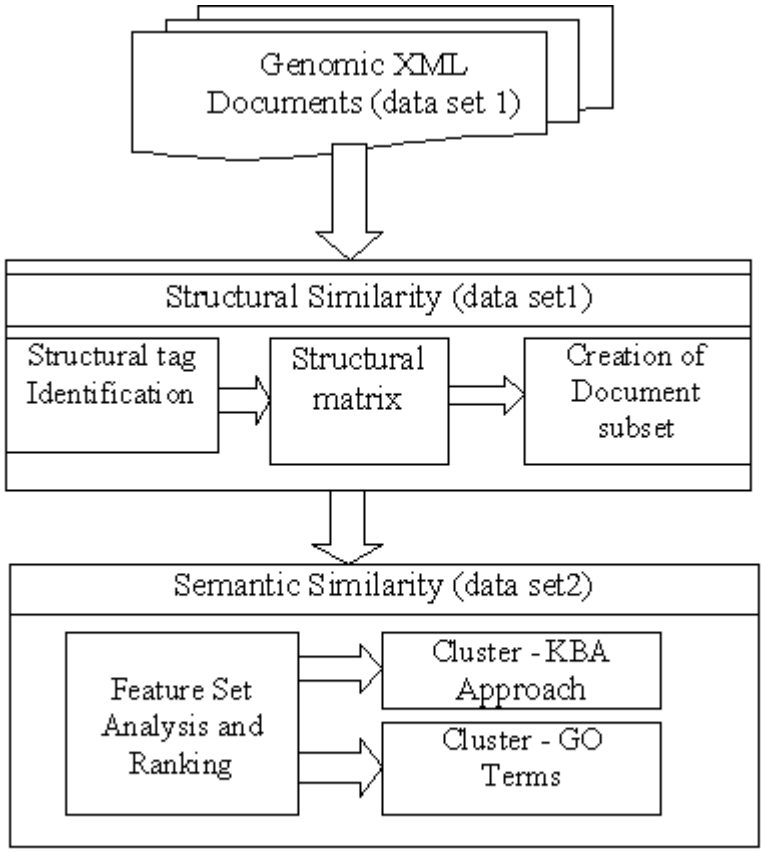
The dataset with 150 documents is given as input for the first phase of clustering to extract documents that are structurally similar which is heterogeneous containing information for two organism E.Coli and human. The structural path is used to analyze the structural similarity. Structurally similar 107 documents were retrieved based on the approach, which is given as input for the content similarity phase.

The various feature sets identified using Filter based on the r anking of rx 2v statistics is shown in Figure **??**. We have considered the feature set with count of keywords with high , low and average 156, 77, and 56 respectively for the proposed study.. The identified feature set is passed for finding the semantic similarity using All keyword based and GO Term based approach. The grouping of documents is based on the 156 keywords which are functionally related with each other. The snapshot of the document grouped in some clusters for a set of keywords for the above kernals is shown in Table 1 and Table 2 It is found that same documents are found

in clusters for the kernel values 30 and 50. It is also found the grouping of documents for clusters {10, 15, 19, 19} are different and contains only one document. In order to asses the semantic meaning of the clusters formed biologically we have analyzed the terms related to the clustered documents. We have inferred from our results that the terms that grouped the documents are biologically associated with each other. The terms responsible for one functionality had other related associated terms. The document with keyword cytoplasm had its associated terms like nucleus, cytoskeleton which are called as cellular components which is associated with a gene name, and Go number. The documents with term oxidation reduction had related terms like fatty acid metabolism, biosynthetic receptor which is responsible for biological activity. The terms like Aledhydde dehydrogenase is associated with keywords like lipid binding, protein binding etc. The above inference of our results motivated us to go for the second approach proposed to cluster documents based on the Go terms and genes which is used by many researchers for gene clustering. The results of our second approach are briefed below b) Go Term Based Approach -Gota Clustering documents based on functionality of the genes using GO terms is proposed in the second approach with the same dataset. The gene names and the corresponding go terms are extracted for the documents which are structurally similar. A total of 71 gene names and 238 go terms were extracted from the dataset and stored in structure for further processing. On implementation of the clustering algorithm the documents were grouped into 30 clusters. The number of documents grouped in each cluster is given in Figure 4. The two approaches Keyword Based and GO Terms based cluster results are shown in Figure 5. The go term 55114 grouped 54 documents which is responsible for the biological activity the oxidation reduction and GO Term 5737 is responsible for cellular activity in cytoplasm. The term 5488 is responsible for Molecular function for binding. The clusters with Go Terms are highly semantically relevant based on functionality than keyword based approach. Some of the documents were found to be overlapped because the functionality of one process inhibits the other. The goterms and its associated genes are functionally related to a process which can be found in the Go Taxanomny. From the results we state that the GO annotations is remarkably useful for grouping documents based on the functionality rather than using the conventional methods The GO Terms and its associated genes are functionally related to a process which can be found in the Go Taxanomy. From the results we state that the GO annotations is remarkably useful for grouping documents based on the functionality rather than Considering other features. The experimental results it is found that the GO terms 55114 grouped larger documents in the second approach which is responsible for oxidation reduction. The same keyword grouped documents for kernel 45 in all KBA. The documents also found distributed in the remaining clusters of the first approach , based on the specific keywords , However the biological inference of both approaches are similar , based on the literature. This paper presents an approach to cluster xml genomic documents using both syntactic and semantic approaches. The structural similarity of documents is done based on the path similarity as in xml documents all information is maintained in tags. The dataset used in the work contains heterogeneous documents with different structural tags for different taxonomies. The structurally similar documents filtered are analysed for Features set Identification using Filter Based approach. The attributes were statistically analyzed and identified three best feature sets. The feature set is used for grouping documents using two proposed approaches Keyword Based Approach and Go Term based approach. The two approaches are compared for their biological relevance. The experimental results it was found that GO Term based clusters documents based on functionality and the terms are related with keywords. Finally, we conclude that using the GO annotations as feature set is efficient to cluster documents which also reduce the dimension of the datasets. [1]

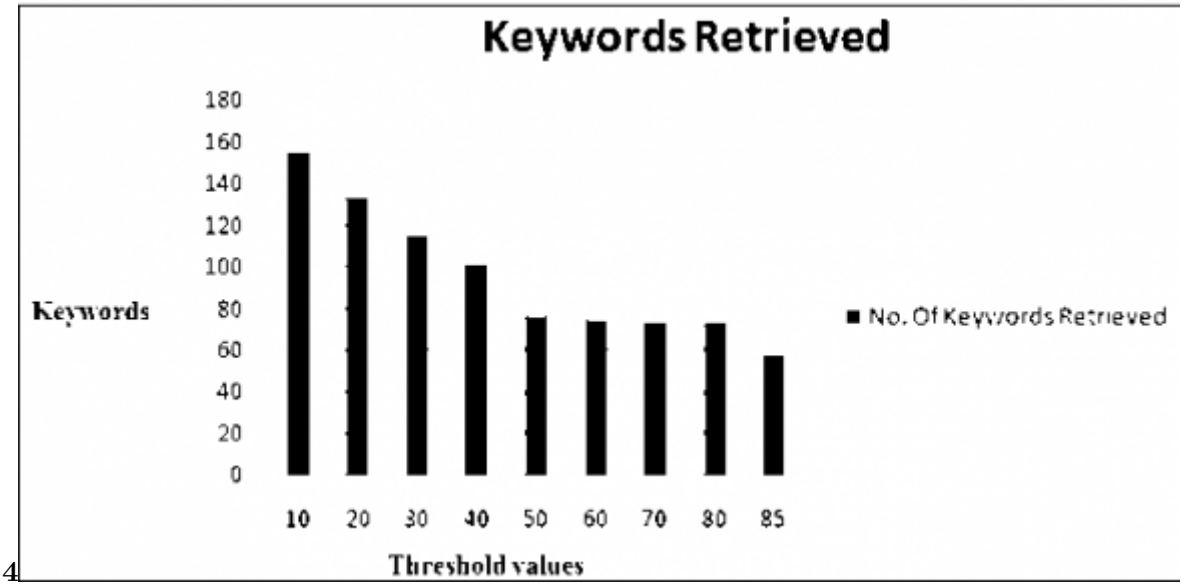
1

Figure 1: Fig. 1 :



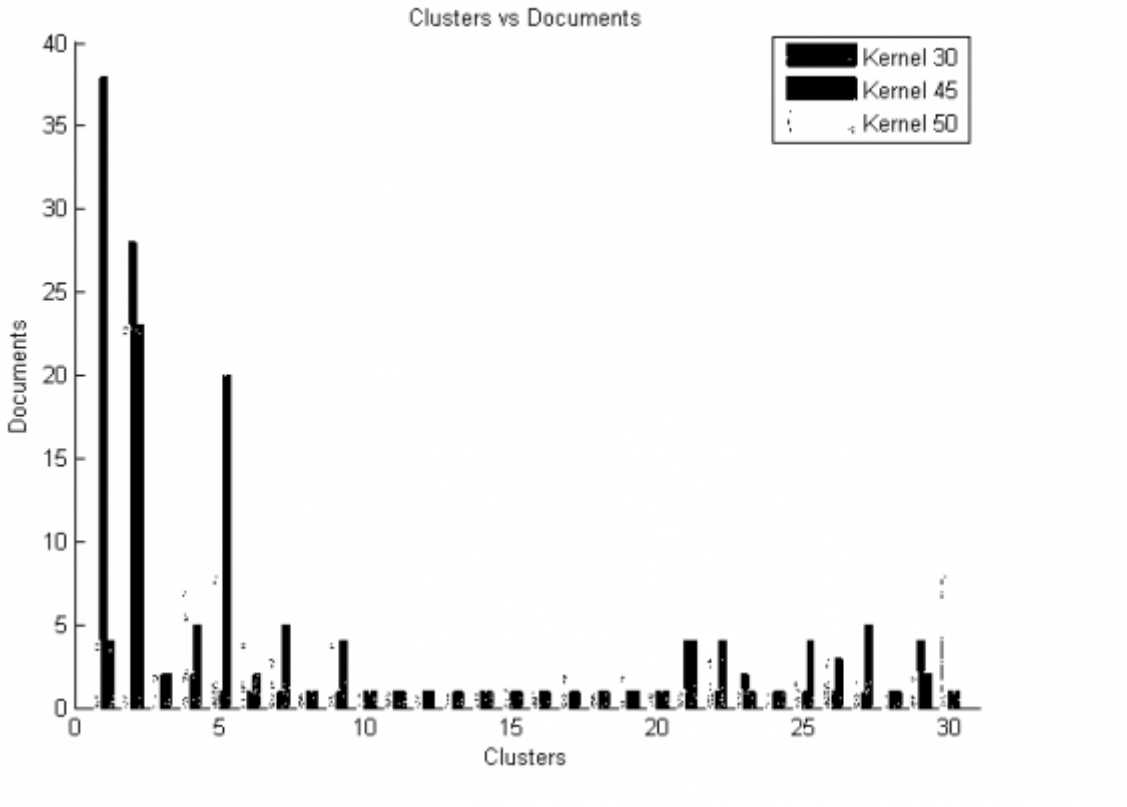
23

Figure 2: Figure 2 :Figure 3 :



4

Figure 3: Figure 4 :



5

Figure 4: Figure 5 :

1

Figure 5: Table 1 :

2

Figure 6: Table 2 :

This work was performed as part of the Minor Research Project, which is supported and funded by University Grants Commission, New Delhi, India.

[Scalable Algorithm for Clustering XML Documents] , *Scalable Algorithm for Clustering XML Documents*

[Bertino et al. (2004)] 'A Matching Algorithm for Measuring the Structural Similarity between An XML Document and A DTD and its Applications'. Elisa Bertino , Giovanna Guerrini , Marco Mesoti . *Information Systems* March 2004. 29 (1) .

[Schlicker et al. (2006)] 'A new measure for functional similarit of gene based on Gene Ontology'. Andreas Schlicker , S Francisco , Domingues . *BMC Bioinformatics* June 2006.

[Sheehan and Quigley (2008)] 'A relational based measure of semantic similarity for gene ontology'. Brendan Sheehan , Aaron Quigley . *BMC Bioinformatics* Nov. 2008.

[Chabalier and Mosser (2007)] 'A transversal approach to predict gene product networks from ontology-based similarity'. Julie Chabalier , Jean Mosser . *BMC Bioinformatics* July 2007.

[Wang et al. (2004)] 'An Efficient and by Structure'. Lian Wang , David Wai-Lok Cheung , Nikos Mamoulis , Siu-Ming Yiu . *IEEE Transactions on Knowledge and Data Engineering* January 2004. 16 (1) .

[Dalamagas et al.] 'An Evaluation on Feature Selection for Text Clustering'. Theodore Dalamagas , Tao Cheng , ; Tao Liu , Shengping Liu , Zheng Chen . *Proceedings of the Twentieth International Conference On Machine Learning(ICML-2003)*, (the Twentieth International Conference On Machine Learning(ICML-2003)) (Clustering XML Documents by Structure" 17)

[Ma and Chbeir ()] 'Content and Structure Based Approach for XML Similarity'. Y Ma , R Chbeir . *Proceedings of the 2005 Conference on Instructional Technologies (CIT "05)*, (the 2005 Conference on Instructional Technologies (CIT "05)Binghamton, Canada) 2005. 2005. p. .

[Jagadasih (2002)] 'Evaluating structural similarity in XML documents'. Nierman Jagadasih , H . *proceedings of the WebDB Workshop*, (the WebDB WorkshopMadison) June 2002.

[Mistry and Pavlidis (2008)] 'Gene Ontology term overlap as a measure of gene functional similarity'. Meeta Mistry , Paul Pavlidis . *BMC Bioinformatics* Aug 2008.

[Pesquita and Faria (2008)] 'Metrics for Go based protein semantic similarity: a systematic evaluation'. Catia Pesquita , Daniel Faria . *BMC Bioinformatics* April 2008.

[Viglianti (2007)] 'MusicXML: An XML based approach to automatic musicological analysis'. Raffaele Viglianti . *Conference Abstracts of the Digital Humanities 2007 conference*, (Urbana-Champaign, Illinois, USA) Jun. 4-8 2007. p. .

[Doucet and Ahonen -Myka ()] 'Naïve Clustering of a large XML Document Collection'. A Doucet , H Ahonen -Myka . *Proceedings of the 2002 Initiative for the Evaluation of XML Retrieval Workshop*, (the 2002 Initiative for the Evaluation of XML Retrieval Workshop) 2002. p. .

[Shen et al.] Yu-Chih Shen , Jia-Lein Hsu , Shuk-Chun Chung . *MF Tree: Extracting and Clustering the Structural Features from Music Object ib MusicXML*,

[Al-Salemi and Aziz ()] 'Statistical Bayesian Learning for Automatic Arabic Text Categorization'. Bassam Al-Salemi , Mohd. Juzaiddin Ab Aziz . *In Journal of Computer Science* 2011. 7 (1) p. .

[Tagarelli et al. ()] A Tagarelli , S Greco , Semantic , Clustering . *Proceedings of the 2006 Siam Conference on Data Mining (SDM "06)*, (the 2006 Siam Conference on Data Mining (SDM "06)Maryland, USA) 2006. 2006. p. .

[Ganseman et al.] 'Using XQuery on Musical Databases for Musicological Analysis'. Joachim Ganseman , Paul Scheunders , D" Wim , Haes . *Proceedings of ISMIR 2008 -Data Exhange*, (ISMIR 2008 -Data Exhange)

[Nayak and Xu ()] 'XCLS: A Fast and Effective Clustering Algorithm for Heterogeneous XML Documents'. R Nayak , S Xu . *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD "06) (The Singapore*, (the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD "06) (The Singapore) April 9-12, 2006. 2006. p. .

[Antonellis et al. ()] *XEdge : Clustering Homogenous and Heterogeneous XML Documents using Edge Summaries*, Panagiotis Antonellis , Christos Makris , Nikos Tsirakis . March 16-20, 2008. Fortaleza, Brazil: ACM.