Feature Analysis for Semantic Clustering of Sequence Documents

V.Bhuvaneswari
Assistant Professor
Department of Computer Applications
Bharathiar University, Coimbatore

Dr.B.LShivakumar
Professor and Head
Department of Computer Applications
SNR Sons College ,Coimbatore

ABSTRACT

The sequence data maintained in public databases are available in heterogeneous formats like FASTA, XML, and ASN1. The XML representation of data is heterogeneous in nature with different DTD in various databases. The difference lies in the representation of sequence description as XML tags. The protein and genomic data in XML format in Genbank has more than 3500 tags to represent the functional description. The sequence documents extracted in any available format, has very vast information related to sequences. Each sequence data has information like its description, alternate names, gene-id, object-id, length, taxon, database references, sequence length and soon. Analyzing the sequence description for understanding the biological process becomes complex due to large number of attributes. The feature selection methods can be applied to select the relevant attributes for genomic and protein dataset. The focus of the study is to select relevant sequence attributes using feature selection methods and semantically group documents.

Keywords:

Feature analysis, GO Terms, Filter based approach.

I. INTRODUCTION

The Genomic and Protein sequence data are stored in public databases like NCBI, Uniport in various formats. The information related to sequence is represented as attributes. Finding the important attributes for clustering genomic and protein sequence, based on annotation, becomes the challenging task. Feature selection methods can be used to analyse and study the features used for representing sequence information using supervised and unsupervised techniques.The focus of the study is assessing features for clustering genomic and protein sequence documents based on content similarity. Feature selection methods are used for analysing features for sequence documents. The task of feature selection is applied for eliminating *irrelevant* features and *redundant* ones. Irrelevant features usually disturb the learner and degrade the accuracy, while redundant features add to computational cost without bringing in new information. The increased dimensionality of data makes testing and training of general classification method difficult. Mining on the reduced set of attributes reduces computation time and also helps to make the patterns easier to understand [4]. The focus of the study is assessing features for clustering genomic and protein sequence documents based on content similarity.

Various approaches are used for clustering documents based on content similarity. The main objective of the paper is to study and analyse genomic and protein sequence features using filter based supervised feature selection methods. Section 2 presents the literature related to the study. Section 3 presents the methodology for feature selection and clustering . In section 4 the experimental results are discussed and finally section 5 draws the conclusion.

II. LITERATURE REVIEW

Feature selection also known as subset selection or variable selection is a process commonly employed in machine learning to solve the high dimensionality problem. It selects a subset of important features and removes irrelevant, redundant and noisy features for simpler and more concise data representation. The benefits of feature selection are multi-fold. In Data Mining, feature selection is the task where we intend to reduce the dataset dimension by analysing and understanding the impact of its features on a model. This chapter discusses the reviews on existing methods for Feature Selection

Qinghua Huang and Dacheng Tao, [3] discussed the importance of feature selection. The objective of feature selection is to find optimal or suboptimal subsets from the original feature sets for irrelevant features removal, intrinsic class information preservation, and improvement of supervised and unsupervised classification performance of classifiers. The methods of feature selection are grouped into two categories, filter and wrapper methods. The filter methods evaluate relevance of each feature and select the features that can maximize some preset performance measures. They are independent of the subsequent learning algorithms. The wrapper method makes use of predetermined learning algorithms to evaluate the feature subsets.

Thangamani M and Thangaraj P, [4] have proposed semantic clustering and feature selection method to improve clustering with semantic relations of the text documents. They have designed a system to

identify the semantic relations using the ontology. The ontology is used to represent the term and concept relationship. They have analysed the performance, accuracy, and efficiency for term clustering and semantic clustering methods.

Li-Ping Jing, Hou-Kuan Huang et al., [1] describes Feature Selection Method using Vector Space Model to provide a convenient data structure for text categorization. According to the empirical results, they have analyzed the advantages and disadvantages and presented a new TFIDF-based feature selection approach to improve accuracy.

Xing Eric P, Michael I. Jordan et al., [5] successfully applied hybrid feature selection methods using filter and wrapper approaches to classification problems in molecular biology. They have also investigated on regularization methods as an alternative to feature selection, and showed that feature selection methods are preferable for the problem.

Mark Devaney and Ashwin Ram, [2] investigated the potential and similar benefits in an unsupervised learning task using conceptual clustering. The issues raised in feature selection in absence of class labels are discussed. They have provided an implementation of a sequential feature selection algorithm based on an existing conceptual clustering system.

Xiubo Geng, Tie-Yan Liu et al., [6] proposed an optimization method for feature selection and ranking. They have discussed the differences between classification and ranking. Their experimental results prove that the total importance scores of selected features are maximized and at the same time the total similarity scores between the features is minimized.

III. METHODOLOGY

The proposed framework shown in Figure 1 is used for the analysis of features in genomic databases. The framework consists of 3 phases the preprocessing phase, Genomic Feature Analysis phase and Clustering phase.

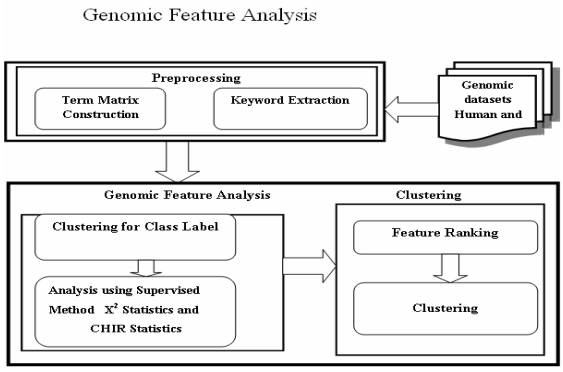

Figure 1 Genomic Feature Analysis

A) Dataset

The *Homo Sapiens* dataset downloaded from the NCBI in XML format is used for the proposed work to analyse and study the sequence features, and cluster similar documents. The NCBI dataset is the integrated, text-based search and retrieval system used at the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others. The dataset consists of 118 nucleotide sequence documents for *Homo Sapiens* taxon.

B)Pre-Processing Phase

The pre-processing phase consists of two important processes. Keyword Extraction and Term matrix construction.

1) Keyword Extraction

In keyword extraction process the attributes are extracted from structurally similar XML documents from DB2 database using XQuery. The key filter interface developed in DB2 XQuery provides the way to extract the necessary fields like gene name and associated go terms, keywords in every biological XML file without converting into any conventional format. We have extracted 396 keywords from *Homo sapiens* dataset from 103 documents. It is identified that some of the keywords were unique for representing particular sequence information. The extracted keywords are used in the next step for term matrix construction. The keywords are analysed to select relevant sequence attributes using filter based approaches.

2) Term Matrix Construction

The term Matrix is constructed after extracting keywords from XML document. The document-term matrix contains rows corresponding to the documents and columns corresponding to the terms. The term matrix is represented in binary encoded format. The value one is entered for the presence of keyword and zero when the keywords does not exists. The XML sequence dataset, have unique domain values for representing functional descriptions string edit measures are not used. The term matrix construction is done using VB.NET and interfaced with MATLAB using COM.

C) GENOMIC FEATURE ANALYSIS

The Genomic Feature Analysis phase consists of two main processes which include:

- Clustering documents initially to assign class label for feature analysis.
- Analysing the features using Filter based feature selection approaches using supervised methods like CHI (χ^2), CHIR.

1) Clustering – Class Label

Document clustering is the act of collecting similar documents into bins, where similarity is some function on a document. The term matrix constructed is given as input for the clustering phase. The documents are initially clustered using hierarchical clustering algorithm for assigning class labels for supervised

techniques. *Homo sapiens* dataset is used for analysing genomic features.

On clustering 103 documents for the datasets 30 clusters are generated to assign class labels. The clusters which contains more than one document is considered to analyse the sequence attributes using filter based supervised method. From the generated cluster the *Homo Sapiens* dataset consists of more than one document in 7 clusters {c2, c3, c17, c26, c27, c29, c30} with a total of 81 documents from which 218 keywords are extracted to evaluate the interdependency of the keywords.

2) Feature Analysis using x^2 Statistics

The supervised feature selection method CHI (x^2) is used to measure the independence between the keyword and the category. From the clustered documents the x^2 values are calculated. There are four steps to find the independency between the keyword and category. First step is to state a hypothesis based on the fit of the data. The null hypothesis and alternative hypothesis are considered. The null hypothesis is that the keyword and category are independent. On the other hand, the alternative hypothesis is that the keyword and category are not independent.

Second step is to build a table of the Observed Frequency and the Expected Frequency. For calculating the x^2, a 2 X 2 contingency table has to be constructed for each cluster with respect to each keyword. The constructed contingency table is said to be observed frequency. The Table 1 shows the general format for the contingency table.

Table 1 Contingency table

	c	$-c$	$\sum$
ω	a	b	$a+b$
ω'	c	d	$c+d$
$\sum$	$a+c$	$b+d$	$a+b+c+d$

- a : denotes the number of documents of category *c* containing keyword *w*.
- b : denotes the number of documents that contain keyword *w* but do not belong to *c*.
- c : denotes the number of documents of category c in which keyword *w* does not occur.
- d : *denotes the number of documents that neither contain keyword w nor belong to c.*

Then the expected frequency is calculated from the observed frequency by using the equation (1).

$$E(i,j) = \frac{\sum_{a \in \{\omega, -\omega\}} O(i,j) \sum_{b \in \{c, -c\}} O(i,j)}{n}$$

…….eq (1)

In the third step x^2 Statistics values are calculated by using the equation 2. It is calculated from the contingency table which includes both the observed and the expected values.

$$x^2(\omega, c) = \sum_{i \in \{\omega, -\omega\}} \sum_{j \in \{c, -c\}} \frac{(O(i,j) - E(i,j))^2}{E(i,j)} \quad eq(2)$$

O (i,j) : denotes the Observed Frequency.
E (i,j) : denotes the Expected Frequency.

In the fourth step, the degrees of freedom for the x^2 Statistics is calculated by using the equation 3.

$$DF = (r-1) * (c-1)$$

Here, r and c are the number of row and column.

In the fifth step, Looking up the x^2 Critical value and the result are concluded. Looking up the x^2 distribution table, if the critical value is much smaller than the x^2 Statistics values, then the null hypothesis is rejected. This can be explained that it is significant and there is some dependency between the keyword and the category.

The Feature set extracted after assigning class labels from the previous phase for the datasets is analysed using the supervised method includes CHI (x^2). 218 keywords for *Homo sapiens* dataset set is used to study and analyse the independency using x^2 statistics. The relationship is analysed between the keyword and category for 81 documents in 7 clusters with 218 keywords.

D) CLUSTERING

The content similarity is the main task involved in document clustering, in which the important terms from the documents that differentiate the documents is to be identified. The identified feature sets from the previous phase is used as input for clustering documents. The Third phase is the clustering phase where the identified feature sets are ranked based on the statistical significance and clustered using hierarchical algorithm.

1) Feature Ranking Based on x^2_{max} , x^2_{avg} and rx^2 Statistics

The Features are Ranked based on $x^2_{max}(\omega)$, $x^2_{avg}(\omega)$ and rx^2 Statistics. The Features with high ranking is used for analysing the term relevance. The best keywords are selected for finding the keyword-goodness. The Supervised feature selection method uses $x^2_{max}(\omega)$ or $x^2_{avg}(\omega)$ to select the best keywords from *m* categories. The x^2_{max} and x^2_{avg} are calculated by using the equation 4 and equation 5.

$$x^2_{max}(\omega) = \max_j \{x^2_{\omega, c_j}\} \quad \ldots\ldots eq(4)$$

$$x^2_{avg}(\omega) = \sum_{j=1}^{m} [p(c_j)] x^2_{\omega, c_j} \quad \ldots\ldots eq(5)$$

Here, $p(c_j)$ is the probability of the documents to be in the category c_j, then keyword whose keyword-goodness measure is lower than a certain threshold would be removed from the feature space. In other words, x^2 selects terms having strong dependency on categories.

The supervised feature selection method CHIR uses rx^2 to measure the Keyword -goodness and prove that rx^2 values represent only the positive keyword-category dependency. The following are the steps to select the *n* keywords. There are three steps to calculate the rx^2 Statistic value. In the first step the rx^2 Statistic value is calculated by using the equation 6.

$$rx^2 = \sum_{j=1}^{m} [p(R_{\omega,c_j})] x^2(\omega c_j, \quad with \quad R_{\omega,c_j} > 1 \qquad eq(6)$$

Here, $[p(R_{\omega,c_j})]$ is the weight of $x^2(\omega c_j,$ in the corpus. In terms of $R(\omega c_j,$, $[p(R_{\omega,c_j})]$ is defined as

$$[p(R_{\omega,c_j})] = \frac{R_{\omega,c_j}}{\sum_{j=1}^{m} R_{\omega,c_j}} \quad with \quad R_{\omega,c_j} > 1 \qquad eq(7)$$

In the second step the keywords are sorted in descending order of their rx^2 Statistic. In third step the top n keywords from the list are selected. The largest values of rx^2 indicates that the keyword ω is more relevant to the category *c*. Keyword with top rx^2 values are chosen as features. The Feature set selected using x^2 Statistics, $CHIR$ Statistics methods are ranked using x^2_{max}, x^2_{avg}, rx^2 Statistic.

The keywords are analysed based on ranking to select the relevant feature attributes. A threshold value is assigned to extract the keywords with good ranking. The Figure 2 shows the threshold value with respect to the keywords retrieved from both the x^2_{max} and rx^2 for datasets.

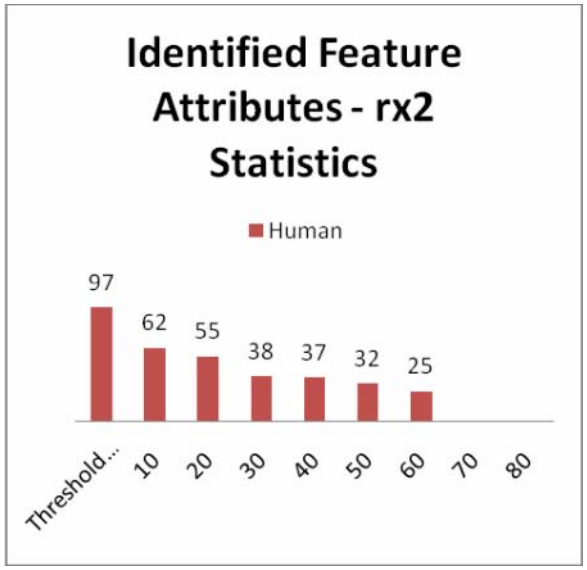

Figure 2 Identified Feature set

2) Clustering – Identified Feature Set

The selected feature set attributes were analysed with respect to the document by varying the no of attributes and clusters were generated. Clustering 103 documents with extracted feature set with 218 keywords we obtained 7 clusters which contain more than one document. Among 7 clusters, 5 clusters {c2, c26, c27, c29, c30} contain maximum number of documents. 81 documents were found to be judged to be of Topic T in entire hierarchy. Remaining documents were found in unique clusters and treated as empty.

In our proposed work the keywords are validated using the F-Measure. F-Measure is the harmonic mean of the precision and recall. Precision is the measure of "exactness". Recall is the measure of "completeness". Precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search, and recall is defined as the number of relevant documents retrieved by the search divided by the total number of existing relevant documents (which should have been retrieved).

$$Precision = \frac{Number\ of\ correct\ answers\ identified\ by\ the\ syste}{Total\ number\ of\ correct\ answers}$$

$$Recall = \frac{Number\ of\ correct\ answers\ identified}{Total\ number\ of\ answers\ specified}$$

The formula for the corresponding Precision, Recall and F-Measure is shown in the eq.12, eq.13, and eq.14 respectively. For any Topic T and cluster X:

$$Precision = \frac{N_1}{N_2} \qquad eq(8)$$

$$Recall = \frac{N_1}{N_3} \qquad eq(9)$$

$$F - measure = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \qquad eq(10)$$

Here, N_1 is the number of documents judged to be of topic T in cluster X, N_2 is the number of documents in cluster X and N_3 is the number of documents to be judged to be of Topic T in entire hierarchy. The selected feature sets are validated by using the above said metrics is discussed in experimental results.

IV. EXPERIMENTAL RESULTS

This chapter discusses and analyses the experimental results of Semantic document Clustering Approaches for genomic and protein sequences using content similarity. The experimental results for supervised feature selection methods x^2 statistics, $CHIR$ Statistics are discussed. This chapter also

discusses the clustering results for grouping documents using the identified feature sets for genomic datasets. The results are verified using F-Measure and validated for biological relevance.

Table 2: Cluster Analysis of Extracted features

F-Measure Analysis - Extracted Feature Set –*Homo sapiens*			
Cluster	Precision (%)	Recall (%)	F-Measure (%)
C2	70	67	68
C26	60	35	44
C27	60	35	44
C29	50	35	25
C30	75	35	47

The Table 1 provides the analysis of clusters before feature selection for the dataset. The F-Measure, Precision, Recall metrics are depicted in Figure 4. The average precision and recall is found to be very low, and many of the documents were also found to be in single clusters. The documents were in single clusters as the key terms of those documents are unique. The feature selection techniques are used to analyse the features and the results are evaluated.

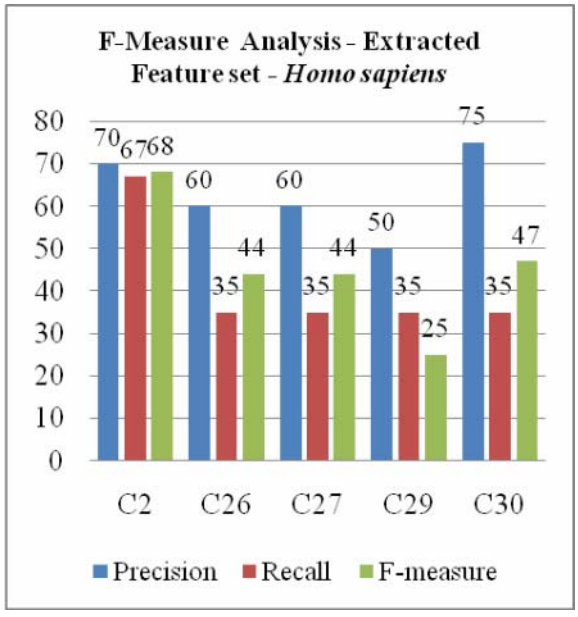

Figure 4 -F-Measure Analysis of Extracted Feature

The Filter based approach is applied and analysing the key terms we have indentified three feature sets based on the ranking of key terms with 97, 38 and 25 keywords for *Homo sapiens* dataset. These feature sets were clustered using hierarchical clustering algorithm and evaluated using F-Measure. The Precision, Recall values is shown in Figure 5. The analysis of Key terms it was found that GO terms dominated in determining the clusters in semantically grouping the documents.

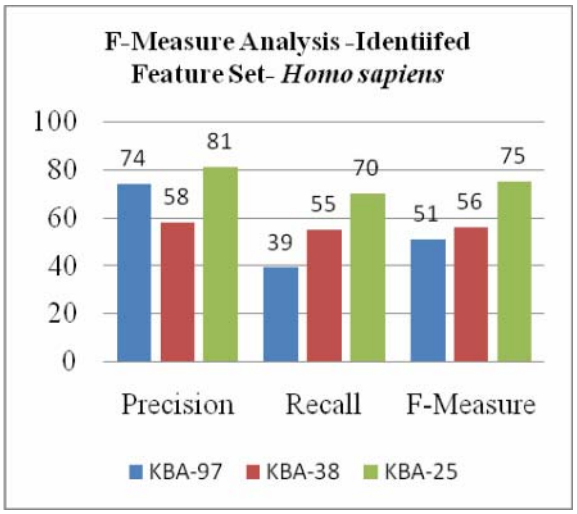

Figure 5 F-Measure for Identified Feature sets

The experimental results of *Homo sapiens* dataset, it is inferred that clusters with GO as key terms groups functionally similar documents than other Keyword with a precision of 81% and recall of 70%. The analysis of biological relevance of GO Term approach, it is inferred clustering semantically similar documents using GO terms is remarkably high and efficient. The documents grouped semantically is high after using the identified feature sets compared to the clustering of the extracted feature set.

V. CONCLUSION

A framework ‘ Genomic Feature Analysis for semantic document clustering for sequence documents is designed and implemented. In first phase, preprocessing the keywords are extracted from XML documents using XQuery to remove noisy data. 396 keywords for *Homo sapiens* dataset were retrieved and term matrix is constructed.

The second phase the Genomic Attribute Analysis, the extracted feature set are analysed and identified using the supervised filter based feature selection methods. Three feature set were identified wit 97, 38 and 25 keywords for *Homo Sapiens* dataset. The identified feature set were analysed for term relevance. In the third phase the identified feature sets are clustered using the hierarchical algorithm and the clusters generated is verified using Precision, Recall and F-Measure.

The experimental results of *Homo sapiens* dataset, it is inferred that using GO as key terms groups

functionally similar documents than other keyword 81% and recall of 70% for feature set with 25 keywords.. The analysis of biological relevance of GO Term approach, it is inferred clustering semantically similar documents using GO terms grouped relevant documents. The identified feature set and can also be used for analysis in document processing.

Acknowledgment

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

References

[1] Li-Ping Jing, Hou-Kuan Huang, Hong-Bo Shi, " Improved Feature Selection approach In Text Mining", In Proceedings of the First International Conference On Machine Learning and Cybernetics, 4-5 November 2002.

[2] Mark Devaney, Ashwin Ram, "Efficient Feature Selection In Conceptual Clustering", Machine Learning: Proceeding of the Fourteenth International Conference, Nashville, TN, July 1997.

[3] Qinghua Huang, Dacheng Tao, "Exploiting Local Coherent Patterns for Unsupervised Feature Ranking", In Proceedings of the IEEE Transactions on Systems, Man, and Cybernetics, 2011, pp. 1083-4419.

[4] Thangamani M, Thangaraj P, "Integrated Clustering And Feature Selection Scheme For Text Documents", In Proceedings of the Journal of Computer Science 6 (5): 536-541, ISSN 1549-3636, 2010.

[5] Xing Eric P, Michael I. Jordan and Richard M. KARP, " Feature selection for high-dimensional genomic microarray data", In: ICML '01: In Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 601–608.

[6] Xiubo Geng, Tie-Yan Liu, Tao Qin, Hang Li, "Feature Selection For Ranking", SIGIR'07, July 23–27, 2007.

Authors

Mrs. V. Bhuvaneswari received M.Phil in Computer Science at Bahrathiar University in 2003, Masters Degree (MCA) in Computer Applications in IGNOU in 2002 and the Bachelors Degree (B.Sc.) in Computer Technology in 1997 from PSG Tech,. She has qualified UGC-JRF in the year 2003. She is pursuing her doctoral research at Bharathiar University in the area of Data mining. Currently she is working as Assistant Professor in department of Computer Applicaion, Bharathair University, Coimbatore, India. Her research interest includes Data Mining, Computational Biology, Bioinformatics and Evolutionary Computing. She has authored more than 20 papers in Journals and Conferences.

B.L.Shivakumar received Ph.D. in Computer Science from Bharathiar University, Coimbatore. M.Phil. in Computer Science from Manonmaniam Sundaranar University, in 2003 and M.Sc. in Computer Science from Bharathidasan University, in 1996. He also received Post Graduate Diploma in Business Administration (PGDBA), Co-operative Management (PGDCM) and Bachelor of Library and Information Science from Annamalai University. In 1997, he joined SNR Sons College as a Lecturer in the department of Computer Science, and currently is the Professor and Head of the department of Computer Applications. He has authored or co-authored over 15 Research Papers in journal and conferences. He is recipient of Bharat Jyoti award conferred by The India International Friendship Society, New Delhi and Best Programme Officer award by Bharathiar University. His interest includes Computer Forensic Science, Digital Image Processing and Cloud computing.