ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 5, Issue 1 (Jan. - Mar. 2017)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2017)

Derived Gene Operational Model for Semantic Similarity Search in Web Document Mining

[I]Dr R.S. Vetrivel, [II]P.M Revathy, [III]P. Jeyanthi Rani
[I]Professor, [II]Research Scholar, [III]Assistant Professor
[I,II,III]Computer Science, Subramanya College of Arts & Science, Palani

Abstract
Web Search is the process of extracting information from World Wide Web (WWW). Document mining research provides high quality information from large collection of database. Relationship among user query and document matching is measured by using similarity scores. Semantic web search provides the information of user's web search queries and different types of web content. Ontology is an essential concept used in the semantic web infrastructure. Ontology collects the search patterns, ideas and contexts in interconnected network.
The existing work introduced a shortest path based on hybrid measure of ontological similarity. It combines structural and semantic information placed in the Gene Ontology (GO) graph. Similarity of the term pair is calculated with the help of weighted path from the lowest common ancestor root. However, the hybrid measure is unable to arrive complete understanding of a biological system. The integration of molecular networks with other data is insufficient. Another one is lack of identifying molecular sequences, protein domains and gene expression profiles.
In order to overcome the above limitations, Derived Gene Operational Model for Semantic Similarity Search in Web Document Mining is designed. The proposed work improved an enhanced gene similarity measure using Derived Gene Operational Model in web document mining. This method develops the effectiveness of transfer functions of semantic similarity of Gene Ontology. The semantic method evaluates the weighted paths for gene ontology similarity measure. Other similarity measures use the derived gene models in their calculation but require the specificity of a concept in hybrid measure. Certain features are used to train the clustering algorithm, in order to classify the web documents.

Key Terms
Gene Ontology, Semantic Similarity, Gene Operational Model, Document Mining

I. Introduction

A. Web Document Mining

Web mining is the important operations of data mining method used to identify the patterns from the Web. Web document mining introduces searching effect of relevant information for retrieving exact users quires and terms. Document data includes text, images, audio, video or structured records and so on. Web document mining technologies are the most important one for extracting knowledge on the Web. Based on the targets analysis, web mining is divided into Web usage mining, Web content mining and Web structure mining. Web content mining is an automatic process that exists over keyword extraction.
Gene Ontology (GO) was created to define the attributes of genes and gene products using a controlled terminology. It is used to support the research related to gene products and documents on the web. The entity type ontology contains the familiar gene ontology of physical types was studied. GO contains collection of defined vocabularies that explain biological models and gene products recognized by computers and individuals. The GO ontology is denoted as a directed acyclic graph (DAG) in which the terms are nodes and the correlation among DAG are edges. DAG is a fundamental structure of the ontology that has child concepts and number of parent.
The several number of parent concepts includes the advantage of higher flexibility, facilitate the powerful grouping, searching and identify the similar genes. Figure 1.1 shows the architecture of gene ontology process. In above entity relationship process, text corpus is taken as input which is accumulated in local database. Initially, the user query is sent to database and related information of the query is extracted from text using the dictionary-based text mining.

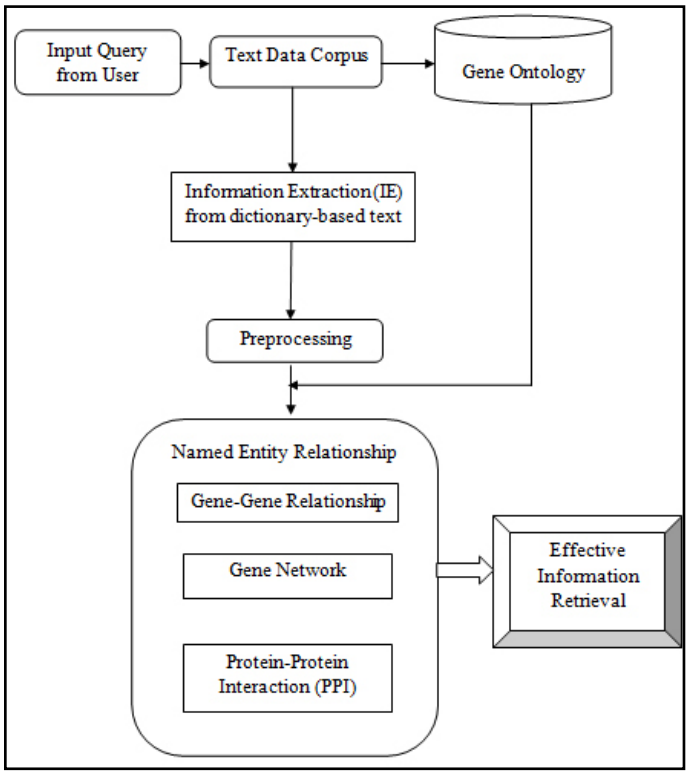

Fig. 1.1 : Architecture Gene Ontology Process

Figure 1.1 illustrates the gene ontology process for retrieving the information about biological data. After the text extraction, the preprocessing is successfully completed to remove the non-functional characters in the extracted information such as stop words, comma, etc. Once the preprocessing step is done then set of the entities are classified in the preprocessed text from

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

text corpus and measure. The semantic similarity involving the terms or entities referred from the above step. After the semantic similarity steps the results produces the entity relationship.
The representation of gene ontology process is widely used for analyzing functionally similar genes and the protein sub cellular or sub nuclear location detection. GO of number biological models are still increased for identifying gene products. GO sets its primary focuses on coordinating this increasing number of models at the risk of losing the characteristics of formal ontology. The development of gene ontology was established to solve this problem and discuss the maintenance of the large-scale biological ontology. Databases and software's are developed and generally available for making it easier to use GO semantic similarity.

B. Basic features of Go Annotations

A GO annotation relates a gene with entities in the ontologies and is produced either by a controller or robotically through predictive methods. Genes are linked with as many terms as suitable as well as with the most specific terms available to reflect newly known about a gene. When a gene is annotated to a term, association between the gene and the terms' parents are successfully recognized.
Because GO annotations to a term, take over the entire assets of the ancestors of the terms. Every path from any term reversed to its root(s) that are biologically accurate or the ontology is revised. Gene annotated to vesicle fusion is recovered based on its entire parent terms for improving the flexibility and power when searching and making inferences about genes.

C. GO and Similarity between Gene Products

The main aim of gene operational model is to describe the similarities among gene products with the help of GO knowledge. Since, each gene product contains number of GO terms, the similarity of gene products determined from the similarities of these GO terms.
The similarities between gene products are calculated by following steps.

- At first, the similarity of two GO terms is identified from the GO graph.
- The next, gene product similarity is obtained from the GO term similarities.

D. Gene Ontology Vs Single-Term Predictions

As a baseline test, the protein function identification is evaluate to operate without GO in place, where the entire association of proteins is analyzed on a single ontology term. The result specifies the power of the network with the construction of Gene Ontology over the single-term network even in the case of number of species networks.
It is significance to specify the model with gene ontology that builds a true positive prediction where the model without assigns a false negative error. This result expected with only one term and one protein annotated to it. In general, incorporating the ontology structure with the requirements of functional terms are efficiently improves the performance over the traditional models.

II. Literature Survey

1. An ontology-based similarity measure for biomedical data – Application to radiology reports

In this paper [1] a semantic vector based method is designed to calculate the similarity among two given documents using systematized nomenclature of medicine -- clinical terms. Semantic based method improves the similarity of documents relating the same analysis. The semantic algorithm develops the classification accuracy when document classification is performed based on imaging process.
Numerous suggestions are introduced to determine the similarity but, there is no standard gold for calculating the similarity. Therefore, a task-based approach is introduced to verify the hypothesized developments in document similarity from the addition of semantics. This effects used for document classification process, which provides its potential application for biomedical information extraction.

2. Detecting Similar Areas of Knowledge Using Semantic and Data Mining Technologies

This paper [2] presents a novel architecture with the help of detecting similar research areas for combining several bibliographic sources. Novel architecture includes the process of extraction, enhancement and characterization of bibliographical resources for detecting patterns using data mining algorithms.
In addition, a centralized repository with bibliographic sources is produced by prototype development. Data mining is applied to determine the similar research areas in Ecuadorian researcher's community.

3. Semantic Similarity Measures in the Biomedical Domain by Leveraging a Web Search Engine

In this paper [3], a page hit counts is planned by applying the Google web search engine for determining semantic similarity between two biomedical concepts. The relevant Google search engine's page counts are collected and calculated for extracting the co-occurrence of two concepts P and Q on the Web.
The similarity scores of multiple patterns are estimated with the help of support vector machines for controlling the strength of semantic similarity measures. The equation F-score is applied for ranking feature extraction. Four kernels of standard support vector classification (C-SVC) are applied for classification models

4. Measure the Semantic Similarity of GO Terms Using Aggregate Information Content

In this paper [4], a novel and successful method is implemented to evaluate the semantic similarity of GO terms precisely and effectively. This method is based on two major observations: (1) In general, the difference of GO terms near the root (more general terms) of GO graph is larger than the terms at a lower level (more specific terms). (2) The semantic context of one GO term is the aggregation of entire semantic values (SVs) of its ancestor terms (include the term itself).
The initial observation replaces the human opinion of term semantic similarity at different domain levels of the ontology. Additionally aggregate information content scheme guarantee the integrity of the semantic information in the semantic similarity measure.

5. An Integrated Approach for Measuring Semantic Similarity between Words and Sentences using Web Search Engine

This paper [5] presents a web-based semantic similarity measure that uses the information exists on the web for determining words and sentence similarity. Web based method develops a page counts and text parts arrived by a web search engine.

Web based measure is generally separated into three types. At first, association rule mining measures rely only on the number of arrived hits. Next, Support Vector Machine (SVM) and integrates both page counts and bits to measure p ranked documents when it is updated. Then the third one measures both combined approaches. Finally, sequential clustering algorithm is designed to estimate both combined (first and second) methods.

6. On retrieving intelligently plagiarized documents using semantic similarity

In this Paper [6], semantic similarity measure uses semantic similarity of words mined inside the data corpus using localized contextual information.

An approach is designed for identifying the plagiarism in text document through semantic similarity measure with Nearest Neighbor (NN) search by kernel in multiclass support vector machine. The approach is tested on plagiarism dataset for increasing the efficiency of solution with different plagiarism level. Semantic kernels identify the plagiarism that outwits with accessible methods. A semantic similarity measure is used for recovering the plagiarized documents. But, the matching efficiency is comparatively lesser.

7. Ontology-based approach for measuring semantic similarity

In this paper [7], semantic similarity between word pairs is essential part of text that allows processing and structuring of textual resources.

The measures are changed to the biomedical field through combining the domain information taken from clinical data or from medical ontologies like MeSH. Information Content (IC) based measures uses the topological parameters of taxonomy to express semantic concept. A new intrinsic IC computing method is designed for depending on taxonomical parameters of ancestors' subgraph. But, the measure failed to identify the similarity measure in efficient manner.

8. An Ontology-Based Semantic Similarity Measure Considering Multi-Inheritance in Biomedicine

This paper [8] joins super concepts of evaluated concepts and specificity feature. The common specificity feature takes depth of Least Common Subsumer (LCS) of two concepts and depth of ontology for attaining the semantic evidence.

The multiple inheritance phenomenons in taxonomy are considered with super concepts. Re-ranking search pages is one of the key issues in IR field. Searching methods depends on keyword matching technique with weaknesses. The web users failed to express the search intention by many keywords. The exactly matched results failed to satisfy the web users. Semantic search engine with page ranking algorithm for finds the data semantically, re-ranks the search results efficiently and place the web results similar for the users.

9. Place recognition based on deep feature and adaptive weighting of similarity matrix

In this paper [9] deep learning and similarity matrix is analyzed for place recognition and infrastructure-free navigation.

For attaining the high representative feature, Convolutional Neural Networks (CNNs) extracts the hierarchical information of objects in image. The image is divided into patches and similarity matrix is created with patch similarities. The overall image similarity is identified through an adaptive weighting scheme with the data difference in similarity matrix. Though, image similarity measurement method failed to calculate the degree of semantic similarity between concepts and words.

10. Clustering clinical models from local electronic health records based on semantic similarity

In this paper [10] key objective is to design the methods for intrinsic similarity-estimation based analysis. For intrinsic similarity estimation, it is depending on established ontology where the SNOMED CT was selected. Lin similarity calculates the similarity together with two aggregation techniques resulting in four methods. The similarity estimations cluster the templates. The test material comprises the templates from Danish and Swedish EHR systems.

11. A semantic similarity measure for linked data: An information content-based approach

This paper [11] presents feature-based definition of Linked Data, a generalized information content-based approach increases the efficiency of existing methods limited specific knowledge representation models.

A document representation method called WordNet-based lexical semantic VSM addresses the existing issues. With help of WordNet, the method used data structure of semantic-element information for classifying the lexical semantic contents and changed EM modeling to disambiguate the word stems. In lexical-semantic space of corpus, lexical-semantic eigenvector of document calculates the weight of every synset. But, it failed to classify the conceptual documents relationship.

12. Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering

In this paper [12], a fuzzy control genetic algorithm (GA) in conjunction is designed with hybrid semantic similarity measure for document clustering.

Clustering algorithms employs the vector space model (VSM) for classifying the conceptual relationships between removed related terms. Semantic space model (SSM) is used as corpus-based method where less dimensions in SSM collects true relationship between documents. Thesaurus-based method is joined with SSM as hybrid plan to semantic similarity measure. In GA, the equalization between capability join to optimum and capacity to find new solutions that affects success for global optimum.

13. Structural similarity for document image classification and retrieval

In this paper [13], a new approach was described the document image structural similarity for applications of classification and retrieval. A codebook of SURF descriptors is taken out from representative training images. Every document is encoded and form spatial relationships by dividing the image and computing histograms of codewords in all partitions. A random forest classifier is used with features for classification and retrieval.

14. Approximate XML structure validation based on document–grammar tree similarity

In this paper [14], an original method is designed for computing the structural similarity between XML document and XML grammar (DTD or XSD) that choose limitation on presence, repeatability and XML elements/attributes. The designed approach uses idea of

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

tree edit distance with new edit distance recurrence and dedicated algorithms for evaluating XML documents and grammar structures as ordered labeled trees. The designed method executes an exact validation with maximum similarity threshold on results.

15. Context-Based Diversification for Keyword Queries Over XML Data

This paper [15] presents an approach diversify XML keyword search with its dissimilar contexts in XML data. A short, vague keyword query and XML data are found and derive the keyword search candidates of the query by feature selection model. An effective XML keyword search diversification model is developed for calculating the quality of all candidates. Three efficient algorithms calculate generated query candidates with diversified search intentions to identify and to return top-k qualified query candidates same as keyword query with large distinct results.

16. Enhanced Associative Classification of XML Documents Supported by Semantic Concepts

In this paper [16], a new approach is designed based on supervised classification to classify XML documents with help of rule based classifier through enriched structure and content features. The methodology addresses the existing issues through accomplishing the classification with structure and content features. It uses ontological information into structural and content based features from XML documents and changes into transaction formats where FP-growth algorithm creates the association rules. An associative classifier eliminates the irrelevant rules from generated association rule.

III. Methodology

Web search has developed rapidly in both research and practitioner communities. Web mining is the term of using data mining techniques for retrieving valuable information from the World Wide Web documents and services. Application of information retrieved process is enhanced by considering semantic relationship between words. It is a fundamental research area in the domain such as natural language processing, knowledge retrieval, document clustering and classification. Ontology is an essential concept used in the World Wide Web (WWW).

Gene networks signify abstract models for difficult interaction approaches between genes and proteins. The gene ontology consists of a limited vocabulary, annotating a gene or gene product to structure in directed acyclic graph. Gene semantic relations connect the terms that characterizes knowledge of functional description and biological component information of gene products. Gene ontology similarity produces a numerical illustration of biological relationship between a gene set used to realize different biological facts i.e., protein interaction structural similarity gene cluster.

Arithmetic structures are used to predict the expression levels of a miniature set of genes in few well-studied pathways. But, the normal of genetic interactions among genes remains high level unknown. Therefore, various assumptions concerning genetic interactions handled in unsupervised network reconstruction methods failed to generalize the new interaction types or organisms. In addition, a more interactions involving genes are investigated and statistical learning methods are handled to conclude the nature of interactions and also calculate new unobserved interactions. Although, supervised approach of network reconstruction frequently focuses on a particular type of genetic interaction.

In this work presents a Derived Gene Operational Model for Semantic Similarity search in Web Document Mining (DGOMSS). The proposed work improved an enhanced gene similarity measure using Derived Gene Operational Model in web document mining. This method develops the effectiveness of transfer functions of semantic similarity of Gene Ontology. The semantic method evaluates the weighted paths for gene ontology similarity measure. Other similarity measures use the derived gene models in their calculation but require the specificity of a concept in hybrid measure. Certain features are used to train the clustering algorithm, in order to classify the web documents.

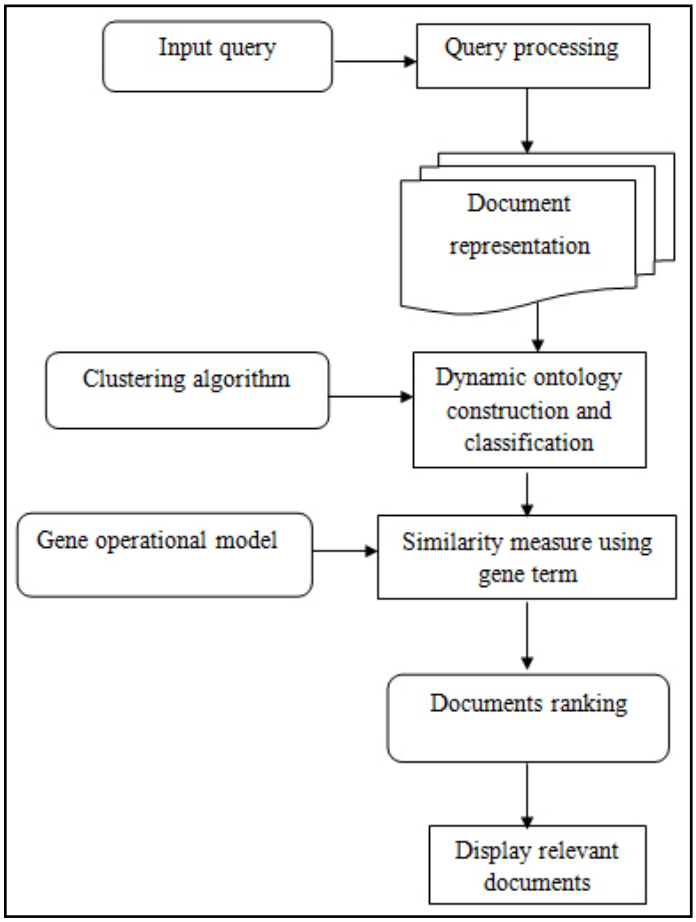

Fig. 3.1 : Architecture for semantic web search using proposed DGOMSS

As shown in figure 3.1, proposed DGOMSS method is divided into several parts such as query processing and documents collection, clustering algorithm, dynamic ontology construction and classification, ranking the documents, etc. Given input user query is first processed to reduce the dimensionality of document representation. After that, documents are selected and classified based on their similarity score to construct ontology. Finally, the classified documents are ranked with the help of ranking algorithm to display the search results based on ranks for user convenience. Proposed gene operational model is applied to retrieve the relevant document in an effective manner. The process of proposed method DGOMSS is elaborately illustrated as follows.

The application of similarity measures over biological gene ontology networks. Clustering algorithms are developed to give superior performance in all these tasks. A lead to suggestion clustering algorithm is primary choice for gene operational model similarity metric over other similarity measures. The derived gene operational model measures for semantic similarity in gene

ontology are divided into three phases:

a) Gene Ontology -document collection
b) Ontology based similarity classification
c) Gene operational model

a) Gene Ontology -document collection

The main objective of ontology construction is to only obtain relevant information when there is large amount of documents by understanding the context of a query. Initially, given input query is processed for providing effective clustering of text documents with the help of clustering algorithm. Biological documents (protein interaction, gene relationships, etc.) are collected from gene ontology where document classification process is to be performed based on given query. Input queries are preprocessed to remove the stop words (for example: is, are, the, they, etc) and hence obtain keywords in a query. Therefore, based on the obtained keywords, relevant information is retrieved from the collected documents. Proper document representation is achieved after preprocessing which helps for providing better results on document retrieval.

b) Ontology based similarity classification

After performing document collection, the next step is to classify the documents with class labels with the help of machine learning algorithm. The similar gene documents are classified from gene ontology. A machine learning algorithm which is employed for documents classification and measuring the similarity between terms. It is trained with partially labeled data obtained from gene ontology. A hyperplane is examined in machine learning algorithm to separate the relevant and other biological documents from the input samples or queries. Hence relevant documents are obtained with the help of classification process.

c) Gene operational model

The derived gene operational model is a discrete genomic centre whose record is regulated by one or more promoters. It contains the information about genes for the identifying the functional proteins or non-coding RNAs and gene products. Semantic relations are used to connect the gene terms for specifying knowledge of functional description and cellular component information of gene products. Similar gene documents (terms) are measured using gene operational model. After the document classification process, the documents are ranked to display the search results. Then relevant document retrieval is done in an effective manner using gene operational model this in turns classification accuracy is improved in proposed derived gene operational model for semantic similarity search in web document mining (DGOMSS) method.

Input: Data set 'D' containing 'n' text documents , Input query with number of terms
Output: relevant document retrieval
Step 1: Begin Step 2: Documents are collected from data set 'D' Step 3: Let each document in 'D' is clustered using clustering algorithm Step 4: Document classification is performed using machine learning algorithm Step 5: Similar gene terms are extracted using gene operational model Step 6: similarity documents are ranked to display search results Step 7: Relevant document is retrieved Step 8: End

Fig. 3.2 : proposed DGOMSS algorithm

IV. Results and Conclusion

A. Performance Metrics

This work quantifies the performance of derived gene operational model for semantic similarity search in web document mining. The analysis result is done through the java platform. This scheme improves the effectiveness of transfer functions of semantic similarity of gene ontology. The performance measures of the proposed work are analyzed with following metrics:

- Similarity ratio
- classification accuracy
- clustering efficiency

1. Similarity ratio

The objective of the validation is to discover whether the estimated semantic similarity is in stroke with the similarity based on the expression data. The similarity based on the gene expression data is computed utilizing the Pearson correlation and is submitted to as the expression similarity. Normally, a high correlation designates a better performance.

Table: 4.1 : Gene Size Vs Similarity ratio

Gene Size	Similarity Ratio (%)	
	HMGO (Existing)	DGOMSS (Proposed)
5	32	38
10	44	50
15	52	58
20	64	70
25	72	78

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

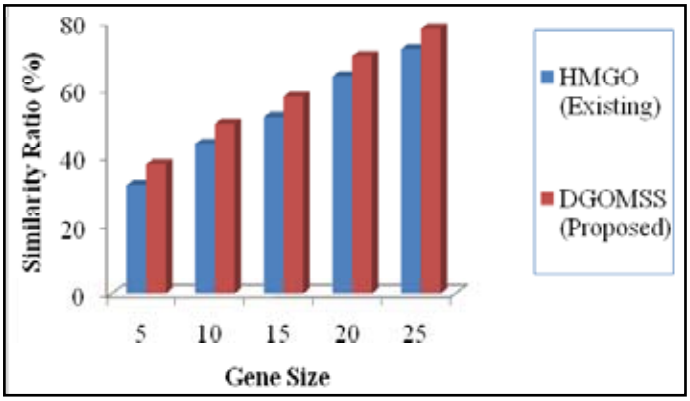

Fig. 4.1 : Gene Size Vs Similarity ratio

Figure 4.1 demonstrates similarity ratio. X axis denotes the gene size values whereas Y axis represents similarity ratio using both the concept of semantic similarity in gene ontology. From table 4.3, the increase in gene size also the similarity gets increased in all methods. The Derived Gene Operational Model for Semantic Similarity (DGOMSS) achieves the high performance of 12 % when compared to the existing system (HMGO).

B. Classification accuracy

Classification accuracy is measured as the ratio of number of correctly classified text documents for a given query to the total number of classified text documents. Classification accuracy is measured in terms of percentage (%). If classification accuracy is high, then the method is said to be more efficient for information retrieval.

Table: 4.2 : Classification accuracy

Number of input queries	Classification accuracy (%)	
	HMGO (Existing)	DGOMSS (Proposed)
10	79	84
20	80	88
30	84	92
40	88	95
50	90	99

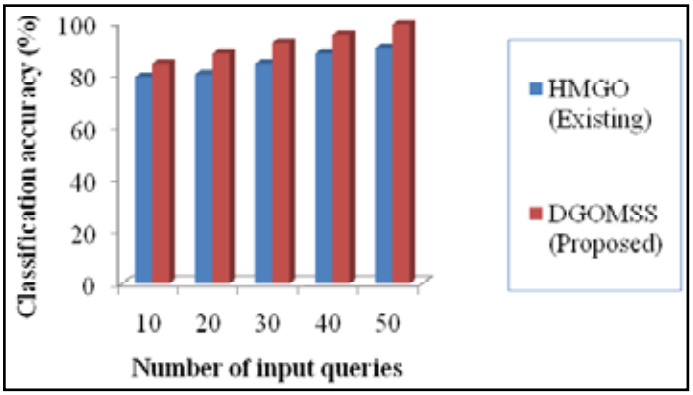

Fig. 4.2 : Classification accuracy

Figure 4.2 demonstrates classification accuracy. X axis represents the number of input quires whereas Y axis denotes classification accuracy using both the concept of semantic similarity in gene ontology. When number of input quires increased, the classification accuracy gets increases accordingly. The Derived Gene Operational Model for Semantic Similarity (DGOMSS) achieves the high classification accuracy of 9% when compared to the existing system (HMGO).

C. Clustering efficiency

Clustering efficiency is measured in proposed DGOMSS method while forming the clusters based on the semantic similarity of given words from different text documents. Clustering efficiency is measured with the help of percentage measure (%). If clustering efficiency is high, then the method is said to be more efficient for further text document classification.

Table: 4.3. Clustering efficiency

Number of input queries	Clustering efficiency (%)	
	HMGO (Existing)	DGOMSS (Proposed)
10	49	54
20	50	58
30	54	72
40	58	75
50	60	79

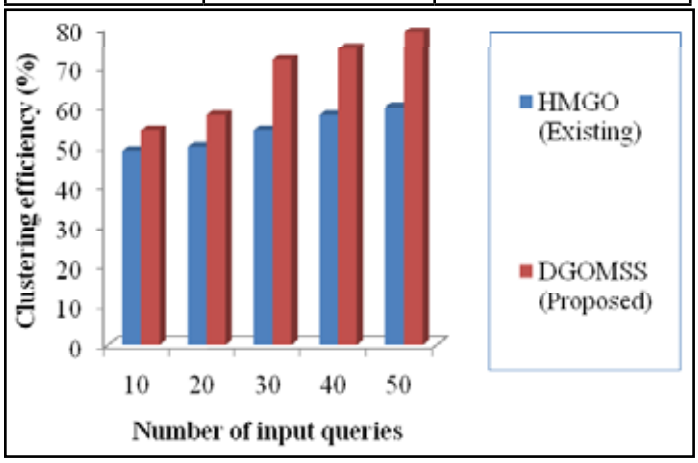

Fig 4.3 : Clustering efficiency

Figure 4.3 demonstrates clustering efficiency. X axis represents the number of input quires whereas Y axis denotes clustering efficiency using both the concept of semantic similarity in gene ontology. When number of input quires increased, the clustering efficiency gets increases accordingly. The Derived Gene Operational Model for Semantic Similarity (DGOMSS) achieves the high clustering efficiency of 24% when compared to the existing system (HMGO).

V. Conclusion

In this paper, a method for measuring the semantic similarity, namely the derived gene operational model for semantic similarity in web document mining is proposed. This scheme improves the effectiveness of transfer functions of semantic similarity of gene ontology. In addition, the document collection process is performed for retrieving the gene ontology terms. Clustering algorithms are developed to give superior performance in all these tasks. In future work, notice that the different type of approach can also be used to improve the classification accuracy for other gene network attributes, such as protein-protein network and the metabolic network.

References

[1] *Thusitha Mabotuwana , Michael C. Lee, Eric V. Cohen-Solal "An ontology-based similarity measure for biomedical data – Application to radiology reports", Elsevier, Volume 46, July 11, 2013, Pages 858-868.*

[2] *Xavier Sumba, Freddy Sumba, Andres Tello, Fernando Baculimaa, Mauricio Espinoza and Victor Saquicela "Detecting Similar Areas of Knowledge Using Semantic and Data Mining Technologies", Elsevier, Volume 329, December 2016, Pages 149–167.*

[3] *Sheau-Ling Hsieh, Wen-Yung Chang, Chi-Huang Chen, and Yung-Ching Weng "Semantic Similarity Measures in the Biomedical Domain by Leveraging a Web Search Engine", IEEE Journal of Biomedical and Health Informatics, volume 17, Issue 4, July 2013.*

[4] *Xuebo Song, Lin Li, Pradip K. Srimani, Philip S. Yu, and James Z. Wang "Measure the Semantic Similarity of GO Terms Using Aggregate Information Content", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 11, Issue 3, May/June 2014.*

[5] *Kavitha Adhikesavan "An Integrated Approach for Measuring Semantic Similarity between Words and Sentences using Web Search Engine" The International Arab Journal of Information Technology, Volume 12, Issue 6, November 2015.*

[6] *Syed Fawad Hussain, Asif Suryani, "On retrieving intelligently plagiarized documents using semantic similarity", Elsevier, Engineering Applications of Artificial Intelligence, Volume 45, October 2015, Pages 246–258.*

[7] *Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Abdelmajid Ben Hamadou, "Ontology-based approach for measuring semantic similarity", Engineering Applications of Artificial Intelligence, Elsevier, Volume 36, 2014, Pages 238–261*

[8] *Fengqin Yang, Yuanyuan Xing, Hongguang Sun, Tieli Sun, and Siya Chen, "An Ontology-Based Semantic Similarity Measure Considering Multi-Inheritance in Biomedicine", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2015, April 2015, Pages 1-9*

[9] *Qin Li, Ke Li, Xiong You, Shuhui Bu and Zhenbao Liu, "Place recognition based on deep feature and adaptive weighting of similarity matrix", Neurocomputing, Elsevier, Volume 199, 2016, Pages 114–127*

[10] *Kirstine Rosenbeck Goeg, Ronald Cornet and Stig Kjær Andersen, "Clustering clinical models from local electronic health records based on semantic similarity", Journal of Biomedical Informatics, Elsevier, Volume 54, 2015, Pages 294–304*

[11] *Rouzbeh Meymandpour, Joseph G. Davis, "A semantic similarity measure for linked data: An information content-based approach", Elsevier, Knowledge-Based Systems, Volume 109, October 2016, Pages 276–293*

[12] *Wei Song, Jiu Zhen Liang and Soon Cheol Park, "Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering", Information Sciences, Elsevier, Volume 273, 2014, Pages 156–170*

[13] *Jayant Kumar, Peng Ye and David Doermann, "Structural similarity for document image classification and retrieval", Pattern Recognition Letters, Elsevier, Volume 43, 2014, Pages 119–126*

[14] *Joe Tekli, Richard Chbeir, Agma J.M. Traina, Caetano Traina Jr., Renato Fileto, "Approximate XML structure validation based on document–grammar tree similarity", Information Sciences, Elsevier, Volume 295, 2015, Pages 258–302*

[15] *Jianxin Li, Chengfei Liu and Jeffrey Xu Yu, "Context-Based Diversification for Keyword Queries Over XML Data", IEEE Transactions on Knowledge and Data Engineering, Volume 27, Issue 3, March 2015, Pages*

[16] *Thasleena N.T, Varghese S.C, "Enhanced Associative Classification of XML Documents Supported by Semantic Concepts", Elsevier, Procedia Computer Science, Volume 46, 2015, Pages 194 – 201.*