Data Mining Approach for Automatic Discovering Success Factors Relationship Statements in Full Text Articles

Dr. M. Somu[1]
[1]Professor,
Department of Computer Science and Engineering. K.S.R. College of Engineering,
Tiruchengode, India.

P. Akshaya[2], M. Gowtham[3],
R. Jayapradha[4], R. Jeevanraj[5]
[2,3,4,5] UG Students
Department of Computer Science and Engineering. K.S.R. College of Engineering,
Tiruchengode, India.

***Abstract*— the data mining techniques are utilized in the context of Business-to-Business (B2B) for identifying sentences that provide the information regarding success factors and their relationships. In this paper is an applying existing technique, especially data mining to automatically classify relevant sentences describing an influencing relationship between success factors word. On the data extraction part, first step is used to select the optimal data mining workflow for automatic classification of sentences. Using Document clustering frame work, focus on correlations between the documents in the local patches is maximized while the correlations between the documents outside these patches are minimized simultaneously. The existing system unsupervised constraints are automatically derived from a two-sided TF-IDF classification model to represent both document and word constraints. It then used an alternating positive and negative class algorithm to optimize the model. The proposed system adopts both supervised and unsupervised constraints to demonstrate the effectiveness of the proposed algorithm in this framework. The proposed N-Gram algorithm applied for data preprocessing to eliminate duplication and apply semantic similarity between words in the documents. The results of the evaluation demonstrate the superiority of current approach against several existing approaches.**

Keywords*— *TF-IDF Classification, Clustering, Similarity Measuure, Term Matrix, N-gram

I. INTRODUCTION

Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns. It has been used for many years by businesses, scientists and governments to sift through volumes of data such as airline passenger trip records, census data and supermarket scanner data to produce market research reports. (Note, however, that reporting is not always considered to be data mining.) A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. Such data are vulnerable to collinearity because of unknown interrelations.

An unavoidable fact of data mining is that the (sub-)set(s) of data being analyzed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships and behaviors that exist across other parts of the domain.

Pre-process is essential to analyses the multivariate datasets before clustering or data mining. The target set is then cleaned. Cleaning removes the observations with noise and missing data. The clean data are reduced into feature vectors, one vector per observation. A feature vector is a summarized version of the raw data observation.

Currently articles that contain huge number of sentence are taken for analysis work. The sentences are split into word and perform sentiment analysis such as positive and negative words. These words are grouped into number of clusters using clustering technique.

In this novel system, once the sentence is split into number of words, they will consider to analysis work where it is separated into string positive and strong negative words. These words are further classified using n-gram technique.

Based on various distance measures, a few methods have been proposed to handle document clustering. A typical and widely used distance measure is the Euclidean distance. The k-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centers. Since the document space is always of high dimensionality, it is preferable to find a low-dimensional representation of the documents to reduce computation complexity.

Low computation cost is achieved in spectral clustering methods, in which the documents are first projected into a low-dimensional semantic space and then a traditional clustering algorithm is applied to finding document clusters. Latent semantic indexing (LSI) is one of the effective spectral clustering methods, aimed at finding the best subspace approximation to the original document space by minimizing the global reconstruction error (Euclidean distance).

However, because of the high dimensionality of the document space, a certain representation of documents usually resides on a nonlinear manifold embedded in the similarities between the data points. Unfortunately, the Euclidean distance is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents.

An effective document clustering method must be able to find a low-dimensional representation of the documents that can best preserve the similarities between the data points. Locality preserving indexing (LPI) method is a different spectral clustering method based on graph partitioning theory. The LPI method applies a weighted

function to each pairwise distance attempting to focus on capturing the similarity structure, rather than the dissimilarity structure, of the documents. However, it does not overcome the essential limitation of Euclidean distance. Furthermore, the selection of the weighted functions is often a difficult task.

Simultaneously, clustering still requires more robust dissimilarity or similarity measures; This paper is motivated by investigations from the above and similar research findings. It appears that the nature of similarity measure plays a very important role in the success or failure of a clustering method.
The first objective is to derive a novel method for measuring similarity between data objects in sparse and high-dimensional domain, particularly text documents. From the proposed similarity measure, new clustering criterion functions are formulated and respective clustering algorithms are introduced, which are fast and scalable like k-means, but are also capable of providing high-quality and consistent performance.

II.LITRATURE SURVEY

Patricia J. Daugherty, R. Glenn Richey, Anthony S. Roath describes [1] Synergy is anticipated when the independent companies work together to plan and execute supply chain strategies. Collaboration between the companies can facilitate both strategic and operational foci, allowing individual supply chain members to exploit their core competencies. In turn, these individual core competencies can help to strengthen the entire supply chain. Conventional wisdom suggests that all firms involved in collaboration should reap greater benefits from working together. A great deal has been written regarding how to set up collaborative arrangements and what's required if supply chain partners want to integrate their operations for mutual gains. Less attention, however, has been focused on strategic aspects. Collaborative efforts often fail because critical long-term details are overlooked.

Adequate care is not taken to select the right collaborative partners, to match inter-organizational needs and capabilities, and to clearly define standards, metrics, goals, and implementation procedures over a planning horizon of one to five years. If collaboration offers such potentially phenomenal gains, it is not worth a more structured and formalized strategic approach? We found both that question and the seemingly mixed signals about collaboration (i.e., it's great, but not really happening yet) intriguing.

Formalization of strategic collaboration may be the overlooked key to making business-to-business relationships work by focusing inter-organizational efforts to achieve maximum gains. Inter-organizational or cross-enterprise supply chain collaboration focuses on sharing of information, joint development of strategic plans, and synchronizing operations. In effect, collaborating partners seek the benefits of vertical integration without the burden of financial ownership. This does not mean, however, that resources are not required to support collaboration. Greatest success is likely when collaborative partners integrate human, financial, and technical resources to create a better business model. Successful supply chain collaboration results in extended enterprises designed to leverage joint capabilities and resources across the supply chain. Each collaborating partner focuses on its unique competency and, working together, the partners can achieve operational excellence that synergistically creates value.

One of the highest profile types of collaboration is Collaborative Planning, Forecasting, and Replenishment (CPFR). Under CPFR, retailers and suppliers work closely together to match supply and demand, thus avoiding expensive overstocks and image-damaging out-of-stock situations. The intent is to work together to coordinate merchandising decisions. Wal-Mart and P&G have experienced significant success through CPFR by jointly forecasting sales of P&G products at WalMart stores, using the resulting information to plan replenishment accordingly (Attaran, 2004). Considering the volume and range of products involved, these CPFR-related improvements translate to significantly better profit margins.
Collaboration requires that diverse entities work together by sharing processes, technologies and data to try to maximize value for the whole group and their customers. Many feel that, in practice, supply chain collaboration has fallen short of its promise **[2]**.

Aaron M. Cohen and William R. Hersh evaluated the major challenge of biomedical text mining [3] over the next 5–10 years is to make these systems useful to biomedical researchers. This will require enhanced access to full text, better understanding of the feature space of biomedical literature, better methods for measuring the usefulness of systems to users, and continued cooperation with the biomedical research community to ensure that their needs are addressed.

The volume of published biomedical research, and therefore the underlying biomedical knowledge base, is expanding at an increasing rate. While scientific information in general has been growing exponentially for several centuries, the absolute numbers specific to modern medicine are very impressive. The MEDLINE 2004 database contains over 12.5 million records, and the database is currently growing at the rate of 500,000 new citations each year. With such explosive growth, it is extremely challenging to keep up to date with all of the new discoveries and theories even within one's own field of biomedical research.

The goal of biomedical research is to discover knowledge and put it to practical use in the forms of diagnosis, prevention and treatment. Clearly with the current rate of growth in published biomedical research, it becomes increasingly likely that important connections between individual elements of biomedical knowledge that could lead toward practical use are not being recognized because there is no individual in a position to make the necessary connections. Methods must be established to aid researchers and physicians in making more efficient use of the existing research and helping them take this research to the next step along the path to practical application.

While manual curation and indexing can be an aid to researchers searching for appropriate literature, a recent study of the information content of MEDLINE records. In contrast to the tagging approach used by Tanabe and Wilbur, Changet al. created the GAPSCORE system, 18 which assigns a numerical score to each word within a sentence by examining the appearance, morphology and context of the word and then applying a classifier trained on these features. Words with higher scores are more likely to be gene and protein names or symbols. After training on the Yapex corpus, 19 precision, recall and F-score were computed for both the exact matches and 'sloppy' matches (defined as a true positive if any part of gene name is predicted correctly), with the system performing much better with sloppy matches .

Accurate text classification systems can be especially valuable to database curators, who may have to review many documents to find a few that contain the kind of information they are collecting in their database. Because more biomedical information is being created in text form then ever before, and because there are more ongoing database curation efforts to organise this information into coded databases than before, there is a strong need to find useful ways to apply text classification. Synonym and abbreviation extraction: Paralleling the growth of the increase in biomedical literature is the growth in biomedical terminology. Because many biomedical entities have multiple names and abbreviations, it would be advantageous to have an automated means to collect these synonyms and abbreviations to aid users doing literature searches.

Furthermore, other text-mining tasks could be done more efficiently if all of the synonyms and abbreviations for an entity could be mapped to a single term representing the concept. Most of the work in this type of extraction has focused on uncovering gene name synonyms and biomedical term abbreviations. Relationship extraction: The goal of relationship extraction is to detect occurrences of a prespecified type of relationship between a pair of entities of given types. While the type of the entities is usually very specific (eg genes, proteins or drugs), the type of relationship may be very general (eg any biochemical association) or very specific (eg a regulatory relationship). Several approaches to extracting relations of interest have been reported in the literature and are applicable to this work. Manually generated template-based methods use patterns (usually in the form of regular expressions) generated by domain experts to extract concepts connected by a specific relation from text.[4]
Automatic template methods create similar templates automatically by generalising patterns from text surrounding concept pairs known to have the relationship of interest. 44,45 Statistical methods identify relationships by looking for concepts that are found with each other more often than would be predicted by chance. Finally, NLP-based methods perform a substantial amount of sentence parsing to decompose the text into a structure from which relationships can be readily extracted.[5] Hypothesis generation: While relationship extraction focuses on the extraction of relationships between entities explicitly found in the text, hypothesis generation attempts to uncover relationships that are not present in the text but instead are inferred by the presence of other more explicit relationships.

The goal is to uncover previously unrecognized relationships worthy of further investigation. Practically all of the work in hypothesis generation makes use of an idea originated by Swanson in the 1980s called the 'complementary structures in disjoint literatures' (CSD). Swanson realised that large databases of scientific literature could allow discoveries to be made by connecting concepts using logical inference. He proposed a simple 'A influences B, and B influences C, therefore A may influence C' model for detecting instances of CSD that is commonly referred to asSwanson's ABC model.

Minlie Huang, Xiaoyan Zhu, Donald G. Payan, Kunbin Qu and Ming Li [6], they describes a novel and robust approach for extracting protein-protein interactions from the literature. Their method uses a dynamic programming algorithm to compute distinguishing patterns by aligning relevant sentences and key verbs that describe protein interactions. A matching algorithm is designed to extract the interactions between proteins. Recently there are many accomplishments in literature data mining for biology, most of which focus on extracting protein-protein interactions. In this survey, they proposed a novel and surprisingly robust method to discover patterns to extract interactions between proteins. It is based on dynamic programming (DP). In the realm of homology search between protein or DNA sequences, global and local alignment algorithm has been thoroughly researched **[7]**. In this method, by aligning sentences using dynamic programming, the similar parts in sentences could be extracted as patterns.

Jan Czarnecki, Irene Nobeli, Adrian M Smith and Adrian J Shepherd describes an important goal of biological text mining is to extract relationships between named biological and/or medical entities. Until recently, the vast majority of research in this area has concentrated on extracting binary relationships between genes and/or proteins, most notably protein–protein interactions. However, attention is increasingly shifting towards more complex relationships, with a particular focus on biomolecular networks and pathways **[8]**. However, in spite of this new focus on networks and pathways, one of the most important sub-topics the construction and curation of metabolic pathways has largely been ignored. This is in contrast to the protein- and gene-centric focus of recent text-mining research: protein–protein interaction networks, signal transduction pathways, protein metabolism (synthesis, modification and degradation) and regulatory networks.

Multiple entity types and entity mismatch. Whereas protein–protein interaction networks, protein metabolism and signal-transduction pathways concern the entity-type protein, metabolic reactions involve both enzymes and metabolites. Moreover, there is a mismatch between the entities that most taggers address (proteins/genes, small molecules) and the entities that we wish to tag in metabolic pathways (enzymes,

metabolites). Similar problems arise in the context of the extraction of protein–protein interactions because protein/gene taggers almost invariably fail to distinguish between proteins and genes. Only a subset of proteins are enzymes, and whereas the distinctive nomenclature associated with enzyme names may be beneficial to the extraction process (we address this point below), it has been argued that identifying the names of metabolites is more difficult than some other categories of chemical name **[9].**

Shahet al.[10] undertook an analysis of the distribution of protein and gene names in 104 articles, and concluded that the Abstract and Introduction were the best sources of information about entities and their interactions, with the Methods and, to a lesser extent, the Results sections often proving problematic (for example, keywords unique to the Methods section commonly refer to reagents and experimental techniques). In this research presented a simple method for extracting metabolic reactions from free text. We have shown that it successfully extracted a high percentage of reactions for two out of three pathways; the third pathway, dealing with fatty acid metabolism, proved particularly challenging owing to the distinctive way in which reactions are described (for example, in terms o molecular addition).

In so far as comparisons with broadly comparable methods are possible, it appears that approach performs rather well; that, at least, is what brief comparison with the performance of gene/protein interaction extraction methods suggests, with both precision and recall at comparable levels. Given that information about secondary metabolites such as ATP is frequently omitted from source papers, we have focussed on the extraction of primary metabolites, rather than side metabolites, in the evaluations we present here. Clearly, this lack of information about side metabolites in the literature is an obstacle to the fully automated construction of complete metabolic pathways using text-mining methods. However, a more realistic goal for a metabolic text mining system is to support manual curation.

Andrea Franceschini, Damian Szklarczyk and Sune Frankild. The STRING **[11]** aims to provide such a global perspective for as many organisms as feasible. Known and predicted associations are scored and integrated, resulting in comprehensive protein networks covering >1100 organisms. Here, we describe the update to version 9.1 of STRING, introducing several improvements: (i) we extend the automated mining of scientific texts for interaction information, to now also include full-text articles; (ii) we entirely re-designed the algorithm for transferring interactions from one model organism to the other; and (iii) we provide users with statistical information on any functional enrichment observed in their networks.

Highly complex organisms and behaviors can arise from a surprisingly restricted set of existing gene families **[12]** by a tightly regulated network of interactions among the proteins encoded by the genes. This functional web of protein–protein links extends well beyond direct physical interactions only; indeed, physical interactions might also be rather limited, covering perhaps <1% of the theoretically possible interaction space (3). Proteins do not necessarily need to undergo a stable physical interaction to have a specific, functional interplay: they can catalyze subsequent reactions in a metabolic pathway, regulate each other transcriptionally or post-transcriptionally, or jointly contribute to larger, structural assemblies without ever making direct contact. Together with direct, physical interactions, such indirect interactions constitute the larger superset of ‘functional protein–protein associations’ or ‘functional protein linkages’

The new version of STRING features a redesigned text-mining pipeline. We have improved the named entity recognition engine to use custom-made hashing and string-compare functions to comprehensively and efficiently handle orthographic variation related to whether a name is written as one word, two words or with a hyphen. As in the previous versions of STRING, associations between proteins are derived from statistical analysis of co-occurrence in documents and from natural language processing. The latter combines part-of-speech tagging, semantic tagging and a chunking grammar to achieve rule-based extraction of physical and regulatory interactions[13].

Evolutionarily related proteins are known to usually maintain their three-dimensional structure, even when they have become so diverged over time that there is hardly any detectable sequence similarity left between them. Similarly, most protein–protein interaction interfaces remain well-conserved over time, at least for the case of stably bound protein partners located next to each other in protein complexes (38,39). This means that a pair of proteins observed to be stably binding in one organism can be expected to be binding in another organism as well, provided both genes have been retained in both genomes. The term ‘interologs’ was coined for such pairs, a combination of the words ‘interaction’ and ‘ortholog’. Whether this high degree of interaction conservation is true also for other, more indirect or transient types of protein–protein associations is less clear—although at least one such type, namely joint metabolic pathway membership, has also been shown to be generally well-conserved[14].

SunePletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X. Binder [15], present a system for extracting disease–gene associations from biomedical abstracts. The system consists of a highly efficient dictionary-based tagger for named entity recognition of human genes and diseases, which we combine with a scoring scheme that considers co-occurrences both within and between sentences. In this paper, shows that this approach can extract half of all annually curated associations with a false positive rate of only0.16%. Nonetheless, text mining should not stand alone, but be combined with other types of evidence.

AUTOMATIC DISCOVERING SUCCESS FACTORS

The paper is focusing on the data extraction part, particularly, on the very first step of selecting the optimal data mining workflow for automatic classification of sentences. The classification divides sentences into a positive class which is a sentence that contains success factors and describes their relationships, and the negative class which is a sentence that does not contain such information. The proposed system developing an application for recommendations of news articles to the readers of a news portal. The following challenges gave us the motivation to use clustering:

- The number of available articles was large.
- Many articles were added each day.
- Articles corresponding to same news were added from different sources.
- The recommendations had to be generated and updated in real time.

Clustering is a technique for automatically organizing or summarizing a large collection of text. The coclustering examines both document and word relationship at the same time. The document similarity is often determined by word similarity, the semantic relationships between words may affect document clustering results. Moreover, the relationships among vocabularies such as synonyms, antonyms, hypernyms, and hyponyms, may also affect the computation of document similarity. The clustering algorithm is reducing and search documents for recommendations in users have been interest to a few numbers of clusters of documents.

This improved the time efficiency and different from sources documents. The main motivation of this work has been to investigate possibilities for the improvement of the effectiveness of document clustering by finding out the main reasons of ineffectiveness of the already built algorithms and get their solutions by applying the K-Means and Agglomerative Hierarchical Clustering methods.

The **Holdout method** is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximator fits a function using the training set only. Then the function approximator is asked to predict the output values for the data in the testing set (it has never seen these output values before). The errors it makes are accumulated as before to give the mean absolute test set error, which is used to evaluate the model. The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the test set, and thus the evaluation may be significantly different depending on how the division is made.

K-fold cross validation is one way to improve over the holdout method. The data set is divided into *k* subsets, and the holdout method is repeated *k* times. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. Then the average error across all *k* trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times. The variance of the resulting estimate is reduced as *k* is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch *k* times, which means it takes *k* times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set *k* different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

Cross-validation, sometimes called rotation estimation, is a model **validation** technique for assessing how the results of a statistical analysis will generalize to an independent data set.

Term Frequency is used to quantify what a document is about. Term frequency (TF) is used in connection with information retrieval and shows how frequently an expression (term, word) occurs in a document. Term frequency indicates the significance of a particular term within the overall document. This value is often mentioned in the context of inverse document frequency IDF. The calculated TF-IDF indicates the importance of each term to the document it belongs to in a context of the whole document. How many times a given word appears in the document it belongs to is the TF (term frequency) part of TF-IDF. The higher the TF value of a given term to a document is the more important the term is for the document.

A. Term Frequency – Inverse Document Frequency (TF-IDF)

The TF measures how frequently a particular term occurs in a document. It is calculated by the number of times a word appears in a document divided by the total number of words in that document. It is computed as TF***(the) = (Number of times term the 'the' appears in a document) / (Total number of terms in the document)***. The IDF measures the importance of a term. It is calculated by the number of documents in the text database divided by the number of documents where a specific term appears. While computing TF, all the terms are considered equally important. That means, TF counts the term frequency for normal words like "is", "a", "what", etc. Thus we need to know the frequent terms while scaling up the rare ones, by computing the following: ***IDF(the) = log_e(Total number of documents / Number of documents with term 'the' in it)***.

For example, Consider a document containing 1000 words, wherein the word give appears 50 times. The TF for give is then (50 / 1000) = 0.05. Now, assume that, 10 million documents and the word give appears in 1000 of these. Then, the IDF is calculated as log(10,000,000 / 1,000) = 4. The TF-IDF weight is the product of these quantities − 0.05 × 4 = 0.20.

IV EXPERIMENTAL ANALYSIS

Fig 4.1 shows the experimental results for Histogram corpus similarity. The figure contains corpus word and frequency count details are shows.

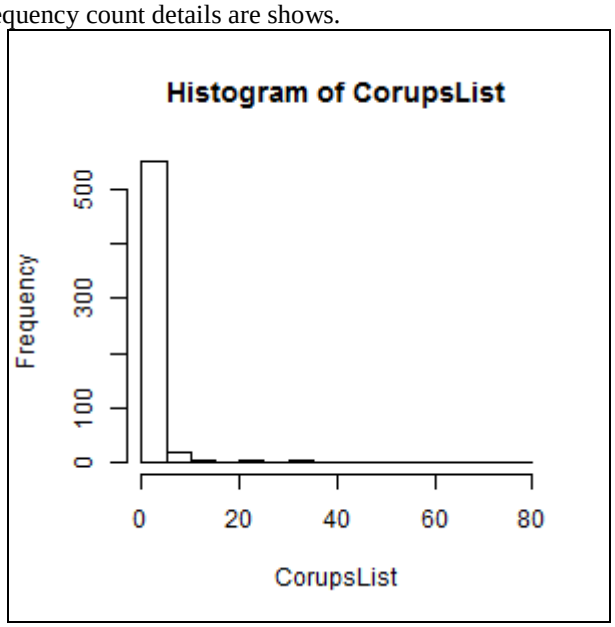

Fig 4.1 Corpus List

Fig 4.2 shows the experimental results for TF-IDF classification similarity. The figure contains corpus word and TF-IDF frequency count details are shows.

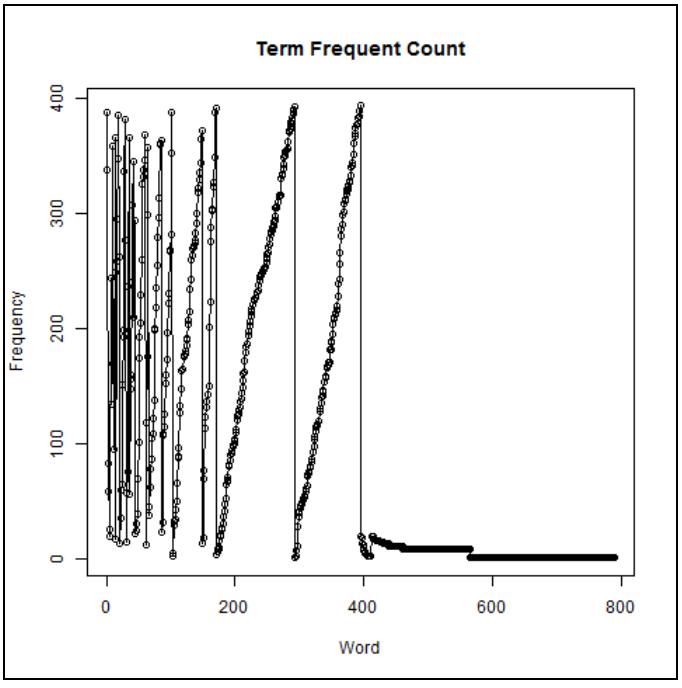

Fig 4.2 TF-IDF Classification

Fig 4.2 shows the experimental results for N-Gram classification similarity. The figure contains corpus word and N-Gram frequency count details are shows.

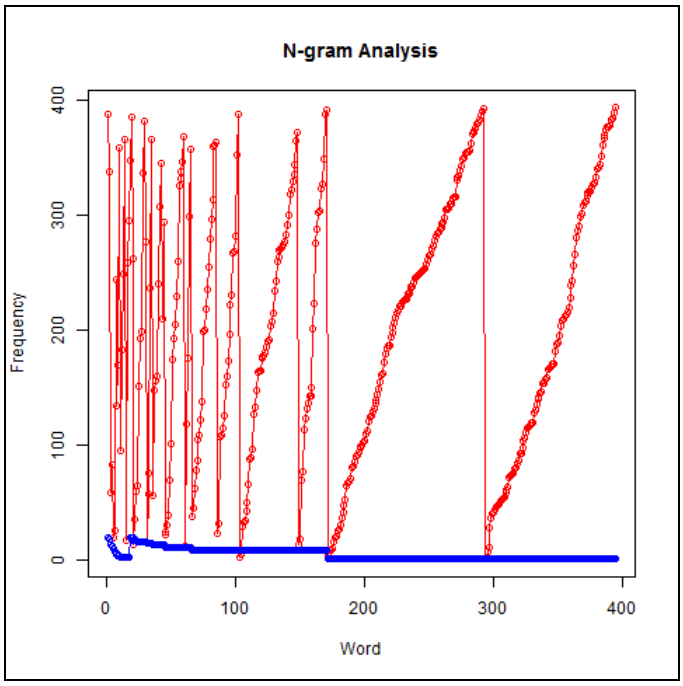

Fig 4.3 N-gram frequency

Fig 4.4 shows the experimental results for Histogram corpus similarity. The figure contains corpus word and frequency count details are shows.

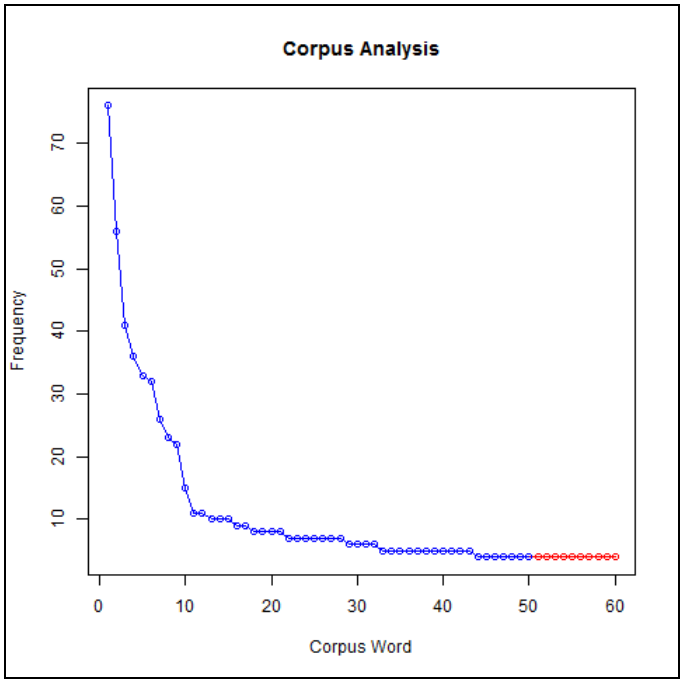

Fig 4.1 Corpus Analysis

V CONCLUSION AND FUTURE ENHANCEMENT

This proposed framework demonstrated how to construct various document and word constraints and apply them to the constrained N-gram process. A novel constrained N-gram approach is proposed that automatically incorporates various word and document constraints into information-theoretic N-gram. It demonstrates the effectiveness of the proposed method for N-gram textual documents. There are several directions for future research. The current investigation of unsupervised constraints is still preliminary. Furthermore, the algorithm consistently outperformed all the tested constrained N-gram with clustering and co-clustering methods under different conditions. The enhanced cosine similarity approach results in better clustering process. The future enhancements can be made for documents of different languages. Investigation for better text features that can be automatically derived by using natural language processing or information extraction tools can be made.

REFERENCES

[1] P. J. Daugherty, R. G. Richey, A. S. Roath, S. Min, H. Chen, A. D. Arndt, and S. E. Genchev, "Is Collaboration Paying Off for Firms?"Business Horizons, vol. 49, no. 1, pp. 61–70, Jan. 2006.

[2] Finley, F., & Srikanth, S. (2005). 7 imperatives for successful collaboration.Supply Chain Management Review,9(1), 30 – 37.

[3] A. M. Cohen and W. R. Hersh, "A Survey of Current Work in Biomedical Text Mining," Briefings in Bioinformatics, vol. 6, no. 1, pp. 57–71, Mar. 2005.

[4] Yu, H., Hatzivassiloglou, V., Friedman, C.et al.(2002), 'Automatic extraction of gene and protein synonyms from MEDLINE and journal articles', in 'Proceedings of the AMIA Symposium', 9th–13th November, San Antonio, TX, pp. 919–923.

[5] Friedman, C., Kra, P., Yu, H.et al.(2001), 'GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles', Bio informatics, Vol. 17, Suppl. 1, pp. S74–82.

[6] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li, "Discovering Patterns to Extract Protein–Protein Interactions from Full Texts," Bioinformatics, vol. 20, no. 18, pp. 3604–3612, Jul. 2004.

[7] Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular sub sequences.J. Mol. Biol., 147, 195-197.

[8] J. Czarnecki, I. Nobeli, A. M. Smith, and A. J. Shepherd, "A TextMining System for Extracting Metabolic Reactions from Full-Text Articles," BMC bioinformatics, vol. 13, no. 1, p. 172, Jul. 2012.

[9] Nobata C, Dobson PD, Iqbal SA, Mendes P, Tsujii J, Kell DB, Ananiadou S:Mining metabolites: extracting the yeast metabolome from the literature. Metabolomics2011,7:94–101. [http://dx.doi.org/10.1007/s11306-010-0251-6].

[10] Shah PK, Perez-Iratxeta C, Bork P, Andrade MA:Information extraction from full text scientific articles: where are the keywords?BMC Bioinformatics2003,4:20. [http://dx.doi.org/10.1186/1471-2105-4-20].

[11] [A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering et al., "STRING v9. 1: Protein-Protein Interaction Networks, with Increased Coverage and Integration," Nucleic acids research, vol. 41, no. D1, pp. D808–D815, Jan. 2013.

[12] Wolf YI, Grishin NV, Koonin EV. Estimating the number of protein folds and families from complete genome data. J.Mol. Biol. 2000;299:897–905.

[13] Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P. Extraction of regulatory gene/protein networks from Medline. Bioinformatics. 2006;22:645–650.

[14] Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2008;36:D623–D631.

[15] S. Pletscher-Frankild, A. Pallej ` a, K. Tsafou, J. X. Binder, and L. J.Jensen, "DISEASES: Text Mining and Data Integration of Disease–Gene Associations," Methods, vol. 74, pp. 83–89, Mar. 2015.