Enhanced Article Discovery Using Document Classification Model

S. Priyanka
Master of Computer Applications
Kongu Engineering College
Perundurai, Erode

L. Rahunathan
Assistant Professor
Department of Computer Applications
Kongu Engineering College
Perundurai, Erode

***Abstract*-The data mining techniques are utilized in day to day application. In this paper, applying existing techniques, presented the experiment method to select a data mining workflow for identifying relevant sentences from full text articles. Based on the preprocessed data, further performed feature selection to extract potential features for sentence classification. The feature selection is performed to obtain the relevant features, using two different approaches such as binary and Term Frequency-Inverse Document Frequency. This paper focus on the data extraction part, selecting the optimal data mining workflow for automatic classification of sentences. The first step sets the term words from the document. Preprocessing techniques are used to normalize the words in each sentence. Then relevant and irrelevant sentences are annotated into strong positive class and strong negative class respectively. using two different approaches Naive Bayes Classifier and Classic K-Means Clustering.**

Keywords-Data mining; naive Bayes classifier; success factor; k-means clustering;

I. INTRODUCTION

Data mining is the process of extracting related data. Data mining is an increasingly essential tool by current business to transform data into an informational advantage. It is currently used in a widespread of reporting practices, such as marketing, fraud detection, and scientific discovery. The related terms data searching, and data interfering refer to the use of data mining techniques to sample portions of the larger population data set that are too small for consistent statistical interpretations to be made about the validity of any decorations discovered. These techniques can however, be used in the creation of new hypothesises to test against the larger data populations.

Performing manual data extraction from text information is timewasting and unpredictable. Document clustering has been used in a number of different areas of text mining and information retrieval. It also has been used to repeatedly generate classified clusters of documents and then uses these clusters to produce an actual document classifier for new documents. Clustering textual data, one of the most important measures is document similarity. Subsequently document comparison is often determined by word similarity, the semantic relations between words may affect document clustering results.

In this paper, we focus on Extracting the text data and arrange them into their corresponding sections. Then, the sentences contained in the text are separated and normalized. Stemming techniques are used to normalize the words in each sentence. This is done to reduce a word. Each word is calculated by considering both the frequency of the word occurring in a sentence as well as the inverse document frequency which indicates how often the word appears across all sentences. The results show that data extraction with clustering and extract only the relevant document based on user query.

This paper is organized as follows: Section 2 provides an overview on related work. In Section 3, the overview of the methodology. Section 4 presents the results of the result and discussion. Finally, an outlook of the conclusion is provided in Section 5.

II.RELATED WORKS

Janos X. Binder, L. Jensen, and S. Pletscher-Frankild [1] describes a system for extracting disease–gene associations from biomedical . 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 takes into account co-occurrences both within and between sentences. In this paper, shows that this approach is able to 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. For this purpose, advanced the illness resource, which assimilates the results from text mining with physically curated disease–gene associations, cancer modification data, and genome-wide association studies from present databases.

Rong Xu, Quanqiu Wang [2] describes a semi-supervised approach to extracting drug-gene relationships from medline. The technique uses one seed pattern and iteratively learns various ways the relationship may be

expressed in medline abstracts. Develop a semi-supervised pattern learning method to extract drug-gene relationships. Personalized medicine is to deliver the right drug to the right patient in the right dose. Pharmacogenomics, the studies in finding genetic variants that may affect drug response, is important for adapted medicine. Computational approaches to studying the relationships between genes and drug response are emerging as an active area of research for personalized medicine. Presently, systematic study of drug-gene relationships is inadequate because a extensive machine reasonable drug-gene relationship knowledge base is problematic to construct and to retain update. The Scientific prose contains rich information on drug-gene relationships, therefore is the vital knowledge source for PGx studies and for adapted medicine. However, this information is largely buried in free text with limited machine understand ability.

J. Czarnecki, I. Nobeli, and A. M. Smith et al. [3] In this paper describes a easy method for removing metabolic reactions from 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 our approach performs rather well; that, at least, is what our 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 focused on the extraction of primary metabolites, rather than side metabolites, in the evaluations we present here. Obviously, this lack of data about side metabolites in the works is an difficulty to the fully automatic creation of complete metabolic pathways using text-mining approaches. However, a more realistic goal for a metabolic text mining system is to support manual curation.

Andrea Franceschini, Damian Szklarczyk and Sune Frankild[4]This paper describe a whole knowledge of all direct and indirect communications between proteins in a given cell would signify an important milestone towards a comprehensive description of cellular mechanisms and functions. Although this goal is still elusive, considerable progress has been made particularly for certain model organisms and functional systems. Presently, protein communications and suggestions are explained at various levels of detail in online resources, alternating from raw data sources to highly formal pathway databases. For many applications, a global view of all the available interaction data is desirable, including lower-quality data and/or computational predictions.

Jason D. M. Rennie, Lawrence Shih, David R. Karger et al. [5] Naive Bayes is often used as a baseline in text classification because it is fast and easy to implement. Its severe assumptions make such efficiency possible but also adversely affect the quality of its results. This paper describes simple, experimental solutions to some of the difficulties with Naive Bayes classifiers, addressing both wide-ranging issues as well as difficulties that rise because text is not actually generated according to a multinomial model. We find that our simple corrections result in a fast algorithm that is competitive with state of-the-art text classification algorithms such as the Support Vector Machine.

M. Huang, X. Zhu, Y. Hao, D. G. Payan et al. [6] In this paper, we propose 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. In our method, by aligning sentences using dynamic programming, the similar parts in sentences could be extracted as patterns. Compared with the previous methods, our proposal is different in the following ways: Firstly, it processes full biomedical texts, rather than only abstracts. Then, it routinely mines verbs for relating protein interactions. Thirdly, this method automatically discovers patterns from a set of sentences whose protein names are identified, rather than manually creating patterns as most previous methods. Lastly, our method has low time complexity. It is able to process very long sentences. In contrast, for any full or partial parsing method, it is time-and memory-consuming to process long sentences.

III. METHODOLOGY

TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY ALGORITHM

Figure 1 describes the overview of method.TF measures how repeatedly a particular word 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. In TF, all the words are considered as important. It counts the term frequency for normal words like “a”, “the”, etc. It is computed by

TF***(the) = (Number of times term the ‘the’ appears in a document) / (Total number of terms in the document)***

IDF measures the importance of a word. It is calculated by the number of documents in the text database divided by the number of documents where a specific word appears. It is computed by

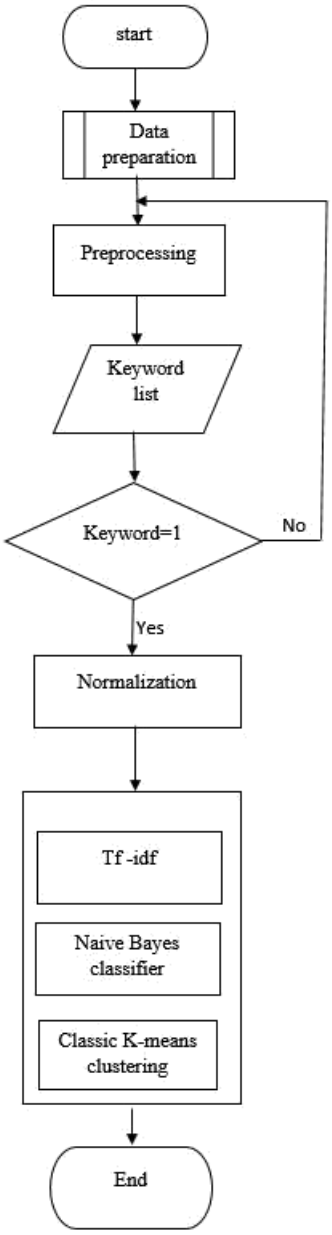

Fig.1: Overview of the Method

IDF(the) = loge (Total number of documents / Number of documents with term 'the' appears)

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.

NAIVE BAYES CLASSIFIER

Naive Bayes Classifier technique is based on Bayesian theorem and it is used for document classification. The document can be classified based on keyword. It is computed by

probability=number of search keyword present/total number of document

In the Bayesian analysis, the final classification is produced based on the probability value.

CLASSIC K-MEANS CLUSTERING

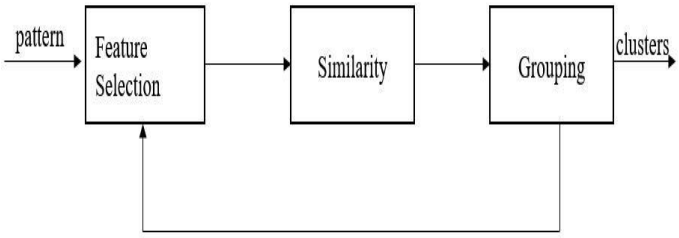

Fig.2: Stages in clustering

Figure 2 describes the stages in clustering.k-means algorithm is one of the easiest clustering techniques used to cluster observations into groups of related observations. The clusters observations into k groups, where k is an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

- The algorithm randomly selects k points as the initial cluster centers
- Each point in the dataset is allotted to the closed cluster, based upon the Euclidean distance between each point
- Each cluster center is recomputed as the average of the points in that cluster
- Steps 2 and 3 repeat until the clusters join. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a difference in the classification of the clusters

- Euclidean Distance: $d(X_i, Xj) = \sqrt{(X_{i,a} - X_{j,a})^2}$

1. procedure KMEANS (X, K)
2. {s1, s2, · · ·, sk} Select Random Seeds (K, X)
3. for i ← 1, K do
4. μ(Ci) ← si
5. end for
6. repeat
7. min k~x n −~μ (C k) k C k = C k [{~x n }
8. for all C k do
9. μ(C k) = 1
10. end for
11. until stopping criterion is met
12. end procedure

IV. RESULT AND DISCUSSION

Table 1 and figure 3 describes experimental result for existing system analysis. The table contains weight of text document, weight of clustering text document and average of text document clustering details are shown.

S.NO	Weight of Document	Weight of Clustering Document	Average of Clustering Document [%]
1	200	155	77.5
2	250	220	88.00
3	300	272	90.66
4	350	322	92.00
5	400	383	95.75
6	450	429	95.33
7	500	468	93.60
8	550	523	95.05
9	600	578	96.33
10	650	633	97.74

Table .1: Average Clustering Documents

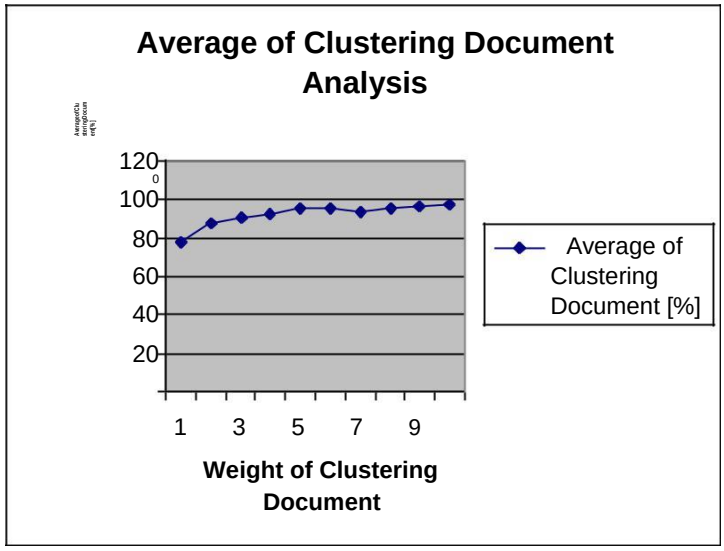

Fig.3: Average Clustering Documents Analysis

V.CONCLUSION

The objective of this paper is to identifying sentences that provide the information regarding success factors and their relationships by utilizing data mining technique. Data extraction with clustering and extract only the relevant data based on user query.

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

VI. REFERENCES

[1] 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.

[2] J. D. Rennie, L. Shih, J. Teevan, D. R. Karger et al., "Tackling the Poor asSumptions of Naive Bayes Text Classifiers," in Proceedings of the 20th International Conference on Machine Learning (ICML-2003), vol. 3. Washington DC, 2003, pp. 616–623.

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

[4] R. Xu and Q. Wang, "A Semi-Supervised Approach to Extract Pharmacogenomics-Specific Drug–Gene Pairs from Biomedical Literature for Personalized Medicine," Journal of biomedical informatics, vol. 46, no. 4, pp. 585–593, Aug. 2013.

[5] 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.

[6] 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].

[7] 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.

[8] Godbole, S., Sarawagi, S., & Chakrabarti, S. (2002). Scaling multi-class Support Vector Machines using interclass confusion. Proceedings of SIGKDD.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10–18, Jun. 2009.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Feb. 2002.

[11] Hahn U, Cohen KB, Garten Y, Shah NH (2012) Mining the pharmacogenomics literature--a survey of the state of the art. Brief Bioinform 13: 460-494.

[12] 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.