Based SVM Distinct Stages Framework Data Mining Technique Approach for Text Extraction

ALSHREEF ABED, JINGLING YUAN, LIN LI
Department of Computer and Science Technology
Wuhan University of Technology
122 Luoshi Road, Wuhan, Hubei, Postcode:430070.
China
lshreefabed2018@hotmail.com; yjl@whut.edu.cn; cathylilin@whut.edu.cn

Abstract: - Today, with high revolution on biomedical domain combined to data mining application, research on text application in the context of information extraction is becoming very essential. Biomedical publications are unstructured in nature, making it difficult to utilize data mining or knowledge discovery techniques to retrieve needed documents. This article introduce a based Support Vector Machine (SVM) Information Extraction from Biomedical domain by adopting the use of text mining framework. Three main structures (Text gathering, Text Pre-processing, Data Analysis) are illustrated to give a input for the text Pre-processing phase. Finally by increasing the precision and recall rate of the information retrieval, good framework performances successfully retrieves the Biomedical documents.

Key-Words: - Data Mining; biomedical text extraction; retrieval; environment; development; information.

1 Introduction

Today research on biomedical domain and its relative areas are growing exponentially very fast. Many modern techniques are introduced with remarkable results, confirming biomedical domain and its use as very potential area of research. The research of biomedical text applied on data mining techniques can provide integral volume of information in the context of text extraction.

If we analyze biomedical based mining only in “extraction” relationship, current works show this position as very complex. It is considered as a key point domain and ultimate because its objective consist in that case to locate the occurrence of a specific relationship type between given two entities. Today, research account varieties of models in knowledge extract developed in biomedical text data. Several of them are known as RDF[1, 2] and model called XML [3, 4, 5]. These two models are totally used today in most of relative articles. For example, in the genomic field, extracting interactions between genes and proteins such as gene-diseases or protein-protein relationships is considered as very essential and are stilling getting deep attraction from international research committee. The context of relation extraction in biomedical text data based mining technique is usually integrated with the similar challenges such as creating high quality annotated data for training or in the context of assessing the performance of relation extraction systems. Among this evolution, text summarization, clustering based approaches, topic modelling, information extraction and clustering technique approaches are represent the most ext mining approaches that currently are used. If we compare relation extraction in the context between some of the different types of biological elements such as genes, proteins and diseases, there are also many progress observed. but in our research, we only focus in the context of system based text gathering, text pre-processing and data analysis.

From the analysis in biomedical relation based “extraction technique”, Knowledge and Information extraction and in particular relation extraction tasks have widely studied various biomedical relations. Many articles using different techniques present biomedical relation extraction with many challenges observed for genes and proteins interactions. Many biological design present different interpretation and most of logical evaluations are also very complex. There are different biomedical technique on biomedical relation extraction but most of them only based on static-syntactic interpretation with and " co-occurrence" focus in most of process illustrated. In this work, the main idea consist to enable data mining and text mining techniques to extract knowledge from such documents and to minimize the number of documents to be checked. To achieve this goal, paper describes a structure called text mining framework which consist to achieve a three distinct stages which include: the Text gathering, the

Text pre-processing and the Data analysis process. The rest of the paper is organized as follows: Section 2 involves the presentation of related works. From Section 3, article provides the insight of the proposed text Mining framework to extract the biomedical documents by adopting the use of the biomedical literature. The experimental setup and results drawn from the proposed work by interpreting the support vector machine (SVM) algorithm are discussed in Section 4. Finally, Section 5 gives the conclusion of the paper

2 Related Works

2.1 Learning Method from Label Data

2.1.1 Feature-based Approaches

There are unlimited list concerning recent researches made on mining technique and knowledge extraction using biomedical text data; the study of biomedical information extraction have been a long process:

In the domain of biomedical information extraction (BioIE), advanced based-featured approaches in the context of engineering have been successfully realized by using machine learning applications. Many relative programs and applications such as dependency tree based features [7], discourse-level features [8], local context [9], lexical, semantic , syntactic, have also been explored. There are other interesting works developed also to improve the performances of Learning from Label Data; several techniques such as combining multiple types of features have been used significantly by combining relatives heterogeneous features, this approach also was another similar approach called lexical, syntactic, semantic and negation features derived approaches from text sentences. Feature text -based Approaches focus more on improving data sparseness efficiency. With the adoption of vector-based word representations and clustering-based word representation, distributional representation new interpretation are performing its general structure. The negative impact on its good ability to provide discriminative power, and one of the remarkable effect observed due to this inability problem is the presence of adverse effects on model learning. The adverse effects on model learning will them increase computational complexity effect and over-fitting phenomenon. As a result, Feature-based techniques are considered as essential for any learning-based approaches, with a big attention especially in the context of high-dimensional features.

2.1.2 Kernel based Approaches

The called “Kernel Approach” is considered also as an essential technique in biomedical information extraction; this technique was very popular during this last five year and more exploited especially in the context of learning algorithms including perception and support vector machines (SVM), although this technique requires high implementation for feature vectors interpretation. One of the best example we can illustrate in this case is the use of sentences which presents more better advantages, this observation is characterized by tree or graph interpretations. We denote also a considerable number of research made in this area during these last few years. In [10], the author presents an approach called " sliding tree kernel " which is a technique based on Kernel applied to perform tree kernel specific. Author adopt the use of a tree's form of sliding structure which is developed realize a consistent model local context of a word structure. In [11] another author present a similar method called hash sub-graph pair-wise (HSP) kernel. This approach was more used in the research deployed in protein-protein interaction extraction, and the main meaning with the use of this technique resides in its capacity to operate in a full dependency graph that represents sentence structure and particularly captures the contiguous topological and label information. Furthermore, in the application of some single-kernel-based approaches, there are possibilities to make outputs fusion when we use different kernel-based systems and multiple kernel learning (MKL) approach in the design of a hybrid kernel. In this case goal focus on linearly including polynomial combining individual kernels interpretations. In [12] author by focusing on DDI extraction introduces a performing system based drug-drug interaction which combined outputs from two kernel-based systems and a case-based reasoning system using majority voting. In the same context, another similar work presented by [13] combined linearly and the shortest path-enclosed tree(SPT) and dependency path tree were extended to capture richer contextual information, this is to achieve good final performance. This new method has been improved by another author by adopting a semantic kernel scheme which is capable to characterize the protein-protein pair similarity. This approach was used on Medical Subject Heading and the context similarity taken into consideration was the Word-Net infrastructure. And a similar work was additionally presented in another relative program called SemEval’13 DDI extraction, which applied an MKL approach linearly combining a feature-based kernel, a shallow linguistic kernel, and

a path-enclosed tree kernel. To resume Kernel Approaches.

There is another work conducted in [14] where authors realized a study review on most 13 popular approaches based kernel type in the context of PPI extraction. The conducted works suggested that the use of the system performance can benefit more from novel features than from novel kernel functions. In the same time it showed a novel framework for biomedical event trigger identification, where word embedding features were combined with syntactic and semantic contextual features using MKL method, achieving the state-of-the-art performance. Analysis showed also improved PPI extraction technique by implementing a model based tree-kernel where processing rules were defined to better handle the parsing error of modal verb phrases and noise interference by appositive dependency.

2.1.3 Cost effective Ground Truth Acquisition

In the context of cost-effective ground Truth acquisition, there is a particular appreciation observed with the use of pre-annotation or computer-aided annotation which is capable to provide human annotators the machine-annotated data with the purpose to realize a potentially better efficiency performance. This method is very helpful and has been approved for creating a dictionary for pre-annotation on clinical NER task where result show a better performance (13-21%) on time needed for biomedical analysis review. Active learning aims to reduce the workload of annotations by reducing the size of the annotation samples. This process chooses the use of informative samples by actively involving learning algorithms. Various studies have investigated active learning in reducing learning costs without affecting the learning performance of related predictive models; with have the case of assertion annotation for medical problems, semantic annotation for medical abbreviations, semantic annotation for medical abbreviations, clinical NER annotation, clinical co-reference resolution, pathological phenomena labeling in MEDLINE, phenotype annotation and other. The use of Cost-effective Ground Truth Acquisition is establishing robustness for extracting medical concepts. The model also significantly reduces the burden of manual annotation.

Crowdsourcing which is another technique has been widely studied in biomedical and clinical fields. Research has shown that this technique is very cheap, fast, and present convenient way to collect high-quality comments in biomedical information extraction. It is important to notice also that a variety of techniques have been explored to improve the quality and effectiveness of crowdsourcing, among that technique, we cite probabilistic reasoning which is used in order to make informed decisions in annotation candidates and game application, and also to motivate the continued participation of experts. Recently, crowd Truth based on crowd commentary result analysis shows that this technique can compensate lack the lack observed in the field of the population's medical expertise. Crowd Truth's experience in the Extract Medical Relations task shows that the crowd and the medical experts can be performed as well in terms of quality and annotating efficiency, and analysis results showed also that each sentence requires at least 10 staff members to receive this comment.

2.1.4 Kernel based Approaches

From the learning in the context of unlabeled data, there is consistent evolution on unsupervised biomedical system developments. Compared to expensive tag data, Biomedical information extraction can get untagged data for free. The main methods include unsupervised, semi-supervised and remote monitoring. The unsupervised biomedical NER system is based on sentence segmentation and distributed semantics, this method shows efficiency the competitive results of clinical notes and biomedical literature. Many research in this context explored core-based pattern and sentence analysis clustering to solve the issue of extracting PPIs and extracting gene-suicide linkages in the biomedical literature. And recently, there is academic works that reported an unsupervised system written in Italian for extracting entities and relationships from clinical records. The semi-supervised approach consist to integrate unlabeled data in a supervisory manner system. The recent researches developed in semi-supervised methods in the context of Biomedical information extraction show that this technique is largely different from the approximate methods used to obtain unlabeled data and uncertainty management when it is added unlabeled data, this approach including several different process such as self-learning for drug discovery extraction, Learning transfer (extracting clinical concepts) and multiple regularization. We have also several strategies used for semi-supervised learning such as active learning with PPI extraction, introducing event inference mechanisms to detect more events recording untagged text and developing topical analysis to determine similar phrases automatically Label.

2.2 Learning from Scheme Integration

2.2.1 Hybrid Approach

In the context of scheme integration, hybrid methods include heuristic/rule/model method, domain knowledge, and learning-based method. In Hybrid methods the overall strategy consists of developing several independent models that work independently and then combining the results of each model to obtain the final result, either by rules or by using a classification/regression model; for example, combining a rule-based model with an SVM classifier for biomedical event detection Pattern recognition is incorporated into learning DDI extraction, and the results of the two methods of algorithm fusion extract temporal relationships in the patient discharge summary. Another general strategy is to run different models in order to further filter and refine them from the system. For example extraction of disease treatment relationships from the MEDLINE corpus or recognition of Genial events by a classifier based on rule-based post processing learning including Similar post-processing have also been used for hybrid biomedical composite feature recognition systems.

2.2.2 Joint Modelling Approach

Biomedical information extraction systems usually involve different subtasks with integrated and interdependent properties. In order to overcome cascading errors in a multi-stage pipeline framework, common models such as the Markov logic network (MLN) method have shown improved performance. Many efforts have been made to alleviate the computational bottleneck of joint inference in the extraction of biomedical events. For example, in [15] several researchers proposed three common models that increase complexity, making the system more robust, and do double decomposition to allow joint inference to be handled; some of them applied a predictive framework based on structured research (SEARN) for high modeling flexibility and rapid joint inference; Venugopal et al. Some of them proposed a MLN-based connection model that uses the SVM model to encode large entities; recently based on the joint decomposition based on the decomposition of the dual decomposition , rich functions based on dependency analysis and word embedding for event extraction have been also proposed. In addition to the above work concerning Biomedical information extraction shared event extraction task, joint modeling has been increasingly applied to other specific sub-domains. To effectively extract adverse drug events (ADEs), many researchers have designed a transformation-based model to jointly extract drugs, diseases, and ADEs that use structured sensor training and multiple beam search algorithms for decoding process.

2.2.3 Open Information Extraction (OpenIE)

The sector of the Open Internet Explorer has been considered as a performed research engine during the last twenty years. It does not work by predefining a predefined set of relationships but aims to identify all possible relationships between untagged and limited data. The Open Internet Explorer system architecture is designed around four main components:(1) Automatic Labeling of data using heuristics or distant supervision [16]; (2) Extractor Learning using relation-independent features on noisy self-labeled data and which represent a database important statistical and analysis of data mining [17]; (3) Tuple Extraction on a large amount of text by the Extractor; (4) Accuracy Assessing by assigning each tuple a probability or confidence score. Based on the functions used in Lara learning, Open Internet Explorer system can divide the existing systems into two categories: mildly open extractors that use only surface language processing, for example. Part of sound marking, and the second category is the segmentation, as well as reopening extractors using deep language analysis. But It has been demonstrated in several types of research that the former is much more efficient, but the recall or accuracy is much lower, while the latter can significantly improve overall performance and affect system efficiency. Open Internet Explorer requires little or no supervision and uses heuristics or remote monitoring to train its extractors for automatic tagging of data, as in traditional boot systems. However, due to its particularly advantageous aspects, it can be well applied to next-generation information retrieval systems; we show in Table 1 below the Advances of Open Internet Explorer system Compared with Traditional self-learning.

One of the big challenge observed with Open Internet Explorer system is that the bulk extraction of Open Internet Explorer systems is purely a textual surface form that cannot be used directly by applications. To avoid this issue in system implementations, there are several ways that can be adopted: (1) the adoption of knowledge resources to understand OpenIE extraction (for example the case of Dynamic Knowledge Graph, Global Network Knowledge Base or the Classification of Text Relation Models); (2) Semantic Web Technology for Information Fusion and Semantic Reasoning; (3) Integration of Ontology Resources, eg Link OpenIE with world knowledge and adjust multiple

ontologies. And in addition, other works also proposed a generic model that links Open Internet Explorer system 's surface shape relationship model with the relationships defined in the knowledge base.

Table 1 Advances of OpenIE Compared with Traditional self-learning

Open Internet Explorer (OpenIE)	Traditional self-learning
Highly scalable to size and diversity of the WEB	Relatively small and homogeneous corpus
Not dependent on relation specific features	Relation dependent features
Avoid lexical features for generalization	Use lexical features for better precision
Domain independent	Domain dependent
No predefined relation schema	Targeted to specific types of relations
Label all the data	Selectively label data
Redundancy-based accuracy assessing	Confidence derived from the learned model

The huge potential of Open Internet Explorer technology for Biomedical information extraction has been officially recognized thank to outperformed approaches introduced during these last twenty years.

3 Proposed Algorithm based Three Distance Stage Framework

3.1 Definition of Algorithm with (SVM) Clusters

First of all, in this approach it is important to notice that in the context of actual information extraction many applicable methods can be used to observe performed result. In this paper context, to provide assistance to the user by showing him the extraction concepts, paper deals with fast clustering concept. Paper adopts the use of Support Vector Machine (SVM) [18],[19],[20] clustering because this method presents non parametric algorithm comparing to Clustering via K-Means approach, SOM-Self Organizing Maps approach and Hierarchical clustering approach frequently used in the same context of implementation.

Gaussian kernel which presents a particularity in width exploration and support vector clustering (SVC) is selected to realize the map of algorithm data points. This operation can presents a high dimensional feature space any times when user is looking for minimal enclosing sphere. The Gaussian kernel use present a width capable to control the scale and if the map back to data space there is possibility to obtain separated algorithm cluster. Two parameters (processing stage and remaining stage) participate to evaluate the structure of a dataset by varying their position with the objective to get a minimal number of support vectors and assure smooth cluster boundaries during extraction process.

3.2 The different Phase Step of Processing

1-Processing: Extraction text processing can transform data into a document matrix model with a frequency indicated for each term. The processing stage which represents the first parameter can provide 70% of extraction work, and the rest of 30% of the work is done by the remaining stage which is the second parameter. The data is prepared for the analysis and that is the first phase, and it represents the most essential phase due to the risk to data to reflect in the remaining phase if the data are not properly pre-processed. The rest of other step includes the

2-Tokenization step: During this phase, the sequence is broken up in strings into data (words, keywords, phrases, symbols). During the process of tokenization, some characters like punctuation marks can be discarded; all sentences are directly converted in raw.

3-The data cleaning: this step consist to consist to remove all previous task and process made previously by other users before getting new specific consigns; during this process all unwanted expressions will be also removed. Some special letter or characters are converted into from upper case letter to low case letter.

4-Stop word removal: all words and expressions that don’t meet the lexical; some expression such as around, at, above, below, bottom, far, many, if …etc are also removed. To avoid the mistake with some biomedical term, some stop words required are collected and listed in a separate file to be upload with the final result.

5-Stemming (Plaice/Husk); in this step, there are different algorithms for rooting identification.

Table 2 Removal Stemming Algorithm with Strength and Similar Affix

Stemmer	Mean Modified Hamming Distance	Median Modified Hamming Distance	Median Characters Removed	Mean Conflation Class Size	Word and Stem Different
Lovins	1.71	0.5	1.66	1.41	3442(68%)
Paice	1.97	1	1.93	1.48	34532(68%)
Porter	1.15	0.5	1.07	1.19	27896(56%)
S-Remove	0.02	0	0.2	1.01	162(3.2%)

According to the new Hamming Distance Descriptive Statics, the result of Paice/Husk algorithm can provide more Mean modified Hamming Distance comparing to other traditional algorithms. This algorithm can eliminate the extreme outliers which contain too many expressions removed.

3.3 Construction of Paice/Husk workflow's Algorithm

Paper designs several rule stem, and most objective with this algorithm is to identify all nous during extraction process. The entire design is resumed in the workflow detailed below Fig. 1:

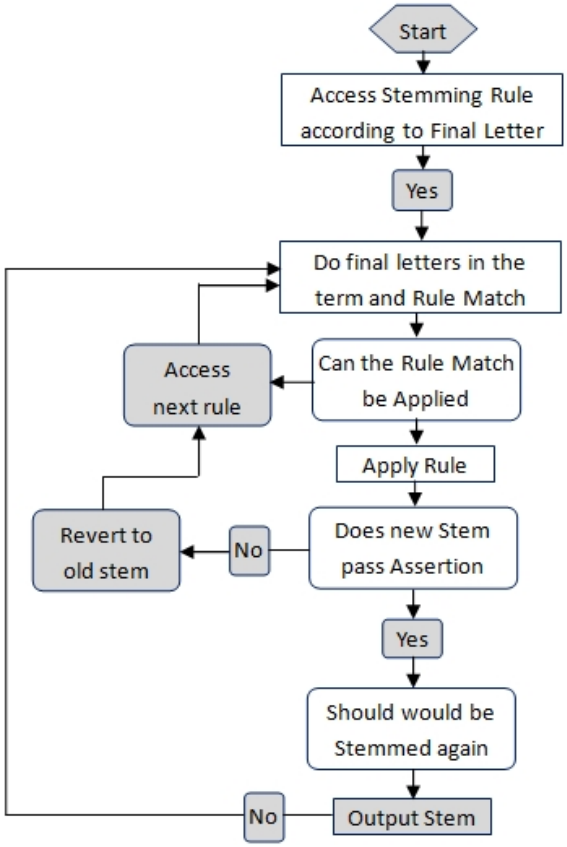

Fig. 1 The workflow of the Paice/Husk algorithm

As we can see from the workflow, the proposed algorithm includes several stem rules the access stemming rule is implemented according to the final letter and all words with for example prefix "ied" should be replaced. If when the system meet this requirement the program confirm by "YES" that the rule is matched and can be applied. If the new result system does not pass the assertion program show "No" and the process to revert old stem will start process.

This technique identified all nouns, verbs and stems and all expressions that grammatically have relative meaning between them are mapped into one cell. The original world obtained after applying the Paice/Husk stemming are listed in the Table 3.

3.4 Description of Support Vector Clustering (SVC) Algorithm

To understand how work the proposed SVC algorithm we description only in this sub-section the Cluster Boundaries process. The Cluster Assignment and the Cluster with/without bounded support vector (BSV) will be directly illustrated in experiment section.

Concerning the support vector machine, the Cluster Assignment consider a denoted $Y_i \leq Z$ the value of the data set representing N points, where $Z \leq IR^d$ the data space. Paper uses a non linear transformation φ from the data space value Z to high dimensional feature-space. The expression below is adopted to realize the enclosing sphere of radius R:

$$||\varphi(Y_j) - C||^2 \leq R^2 \quad (1)$$

Where $||\ .\ ||$ represents the Euclidean norm, C represents the center of sphere. And the Soft constraints are incorporated by adding slack variables represented by the expression δ_j:

$$||\varphi(Y_j) - C||^2 \leq R^2 + \delta_j \quad (2)$$

The super vector clustering SVC satisfied the condition $\delta_j \geq 0$.

Table 3. Paice/**Hush** Stemmer Output

Original Word Abstract with rapid	**Stemmed Word Abstract with rapid**	**Original Word Access Ever-increasing**	**Stemmed Word Access ever-increasing**
Growth	Growth	Quantity	Quantity
Articles	Articles	Information	Information
Genomics	Genomics	Understand	Understand
Research	Research	Rewest	Rewest

The use of Lagrangian is adopted to resolve the problem and the entire expression is illustrated as below:

$$\text{Lag} = R^2 - \sum(R^2 + \delta_j - \|\varphi(Y_j) - a\|^2)\beta_j - \sum\delta_j\gamma_j + \Delta\sum\delta_j \quad (3)$$

Where $\beta_j \geq 0, Y_j \geq 0$ (β_j and Y_j represent Lagrange multipliers), Δ is a constant, $\Delta\sum\delta_j$ represent the penalty term. We set to zero the derivative Lag with respective δ_j,R and a. and we paper evaluate the Karush-Kuhn-Tucker complementarities' conditions of Fletcher such as:

$$\delta_j\gamma_j = 0 \quad (4)$$

$$(R^2 + \delta_j - \|\varphi(Y_j) - a\|^2)\beta_j = 0 \quad (5)$$

Where $a = \sum\beta_j\varphi(Y_j)$, $\beta_j = \Delta - \gamma_j$, by examining Equation (5), approach can determines that the image of Y_j with $\delta_j > 0$ and $\beta_j > 0$ lies outside the feature-space sphere. The Equation states that such a point has $\gamma_j = 0$, so we can affirm that Equation $\beta_j = \Delta - \gamma_j$ detailed in Equation (5) satisfy the condition $\beta_j = \Delta$; This will be called a bounded support vector or (BSV). In this analyze we can affirm that SVs lie on cluster boundaries, BSVs lie outside the boundaries, and all other points lie inside them. Note that when $C > 1$ no BSVs exist because of the constraint $\sum\beta_j = 1$ .Using these relations approach eliminates the variables R, and variable a .Using these relations we may eliminate the variables R, a and γ_j turning the Lagrangian into the Wolfe dual form that is a function of the variables β_j.

4 Simulation Results

4.1 Clustering Dataset and & Bounded Support Vector (BSV) Interpretation

From the analysis of Equation detailed above, BSVs are outside, and all other points lie inside the clusters. The result of Fig.2 is Clustering of a data set containing 183 points which use Support Vector Clustering with $\Delta = 1$. Support vectors are designated by small circles, and cluster assignments are represented by different grey scales of the data points.(a): q = 1 (b): q = 20 (c): q = 24 (d): q = 48.

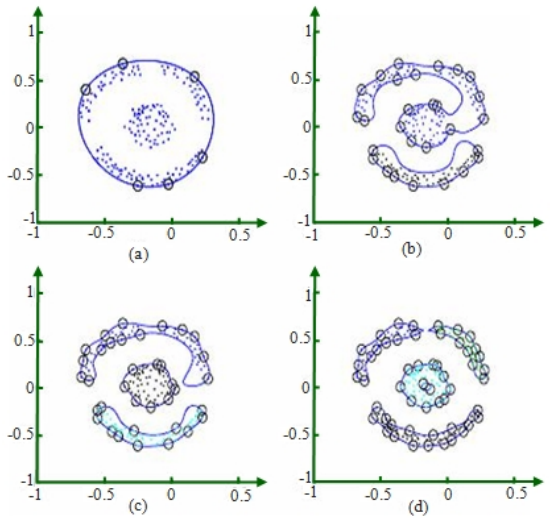

Fig 2. Clustering of a data set containing 183 points.

In the context of clustering assignment, cluster description algorithm does not differentiate between points that belong to different clusters. The adoption of a geometric interpretation based on presentation R(Y) in this case can give a pair of data points that belong to different components (clusters), and all paths that will be connected have to exit from the sphere in feature space. For this reason the segment of each path is represented by a point described by expression $R(Z) > R$.

In the context of clustering with/without bounded support vectors (BSVs), the configuration start with a data set in which the separation into clusters can be achieved without invoking outliers, the paper considers the choose of the constant value determined at $\Delta = 1$. As shown the result in Fig. 3 of the scale parameter of the Gaussian kernel. There are an increasing q value and variance of the shape of the boundary fits more tightly the data. From the Fig. 3a, it can be observed the smoothest cluster

boundary, defined by six support vectors, and by increasing the value q the number of support vector represented by N_{SVs} increase also mutually.

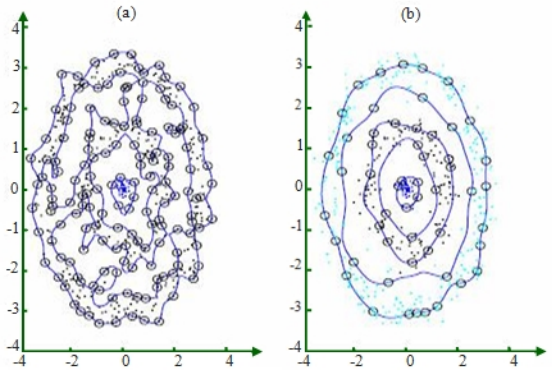

Fig. 3. Clustering with and without BSVs.(a): A ring can not be distinguished when C = 1 here q = 3.4, the lowest q value that leads to separation of the inner cluster (b): Outliers allow easy clustering with parameters p = .3 and q = 1.0 .

4.2 Comparative Study between Biomedical & Other Dataset Scope

In the last part of this experiment results, paper use Distinct stages technique approach to collect and textually analyze several most popular text mining techniques in a comparative approach by selecting more than 400 hundred journal article papers in different research scope and data base from IEEE, Sciences Direct, Cambridge, Google scholar, Evise, Elsevier. The concept of "Biomedical extraction" is adopted to evaluate the best frequency obtained during filtering process. The Table 4 shows the 5 most frequency terms mentioned during data collection from 5 different database. The clouding technique helped to realize the concept and results show that " Biomedical extraction " is the most keyword that was mentioned across all the collected articles. The second highest frequent words are " Patients " and " Students " respectively. The increasing number of the words (learning and students) could be attributed to the fact that learning and students from the core of the higher educational processes. In addition, the appearance of the word (Patients) in some articles it shows that some of contains were focusing on mobile learning in medical education but the high dominance still confirmed by biomedical contains observed in final comparative results.

Paper implement collected word frequency for text extraction as results are shown in Fig.4 the most frequently linked words among Comparative databases: " Learning " followed by " Patients", " Students", " Education", " Care", " Mobile", " Study", " University", " Biomedical", and " Clinical " respectively. These results indicate that the most extracted frequent linked words used in the context of three distinct stages such as text gathering and focus on studies targeting mobile learning in medical education present remarkable performance in the context of extraction medical education and learning.

Table 4 Words Cloud Terms Distribution Across Comparative databases

Database	Term	Frequency
Cambridge	Biomedical	1165
	Education	854
	University	847
	Students	831
	Higher	490
Google scholar	Biomedical	9046
	Care	6458
	Learning	5423
	Medical	5180
	Health	4086
Elsevier	Learning	Biomedical
	Students	1719
	Mobile	1611
	Education	1112
	University	921
Evise	Biomedical	7556
	Patients	6392
	Students	3266
	Care	3212
	Mobile	3193
Science Direct	Biomedical	2043
	Use	837
	Mobile	701
	Education	584
	Student	551
IEEE	Education	793
	Students	777
	Biomedical	579
	Engineering	417
	Higher	256

From Fig. 5, the words concerning by expression" Biomedical " was highly and frequently mentioned by Google Scholar database followed by Evise, Science Direct, Elsevier, IEEE, and Cambridge, respectively. The fig.6 paper presents the terms extracted in the context of "patient" research with high frequency.

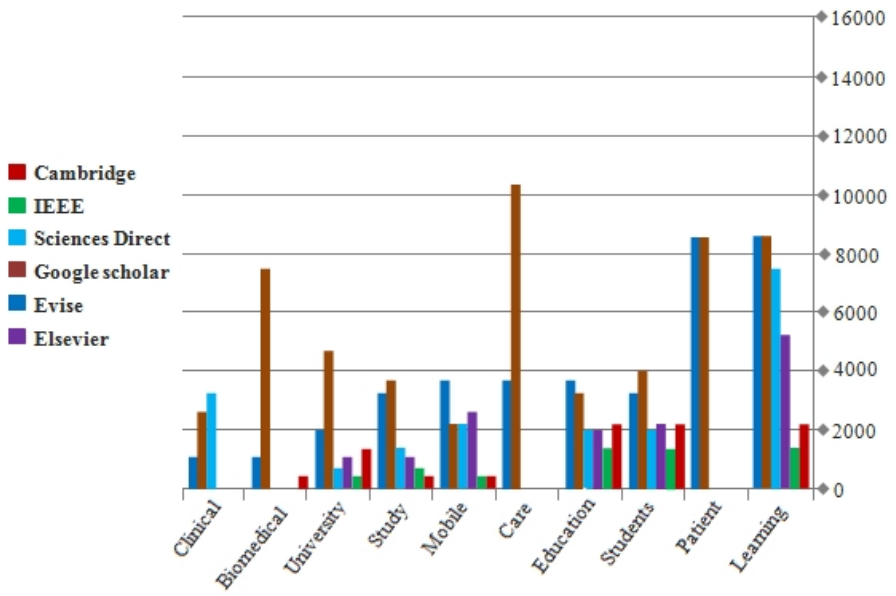

Fig.4 Word Frequency Distribution Across Comparative Results between all Databases

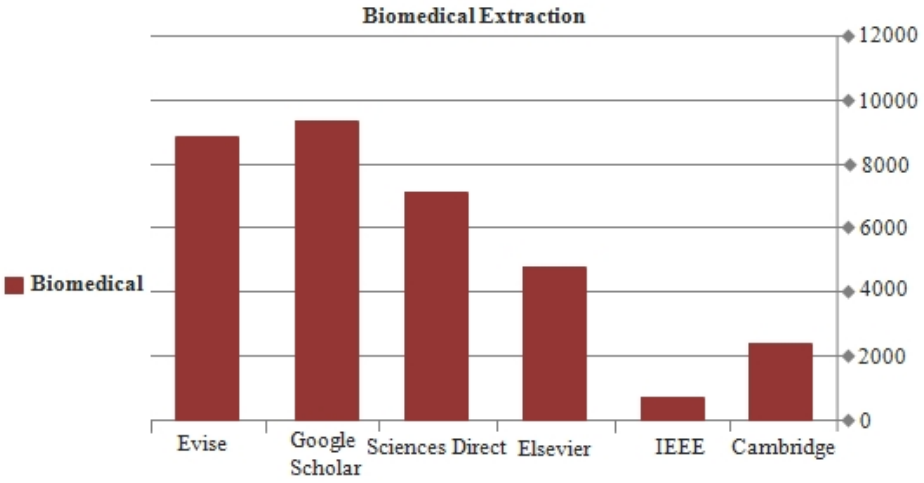

Fig. 5 The Distribution of the Extraction
Result Words in the Field of " Biomedical " among all Sources

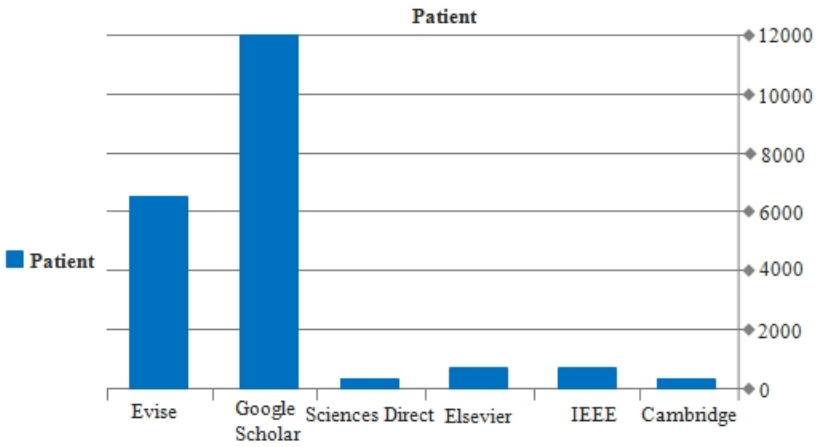

Fig. 6 The Distribution of the Extraction
Result Words in the Field of " Patients " among all Sources

5 Conclusions

In this paper, a proposed based data mining technique approach for text extraction has been developed. Gathering the Biological documents have been extracted from popular Biomedical databases such as PubMed, MedLine, MDB, National Library Of Medicine and MeSH databases. The processed documents are given to the block Data Analysis. The area where the data analyzed by using specialized novel SVM clustering algorithm configured with two points P and q.

By adopting the selection of some patterns such as Patients, Students, Education, Care, Mobile, Study, University, Biomedical and, Clinical to evaluate clustering of the data set and by making comparative approach paper chooses in the comparative context paper uses database from IEEE, Sciences Direct, Cambridge, Google scholar, Evise and Elsevier to evaluate the performance concept of "Biomedical extraction". Three hundred refereed scientific documents from six scientists databases were collected to realize extraction work, and textually analyzed through text mining techniques. The six databases are Science Direct, IEEE, Google Scholar, Evise, Cambridge, and Elsevier. The selection of the collected articles was based on the criteria that all these articles should incorporate biomedical words as the main component in the higher implementation context. In the present experiment, text clustering, association rule, word cloud, and word frequency are the main tasks used for text analysis.

Thus the above framework successfully retrieves the Biomedical documents. The goal of this thesis is to increase the precision and recall rate of the information retrieval. Accordingly, this technique can perceive that information extraction and data mining techniques were never applied to the mobile learning field. This creates a need for collecting a dataset that consists of several research articles in the field of mobile learning. This proposed technique can be used for a variety of research topics, wherein each topic it can generate a wide range of knowledge patterns and particularly in Mobile learning field. because mobile learning has become one of the trendy fields in the higher education.

References:

[1] Carole Goble and Robert Stevens. *State of the nation in data integration for bioinformatics*. Journal of biomedical informatics 41, pp: 687–693, 2018

[2] Andrew Newman, Jane Hunter, Yuan-Fang Li, Chris Bouton, and Melissa Davis. *A scale-out RDF molecule store for distributed processing of biomedical data*. In Semantic Web for Health Care and Life Sciences Workshop, 2008.

[3] Nathan Bales, James Brinkley, E Sally Lee, Shobhit Mathur, Christopher Re, and Dan Suciu. *A framework for XML-based integration of data, visualization and analysis in a biomedical domain*. In International XML Database Symposium. Springer, pp:207–221, 2015.

[4] Mahmood Doroodchi, Azadeh Iranmehr, and Seyed Amin Pouriyeh. 2009. *An investigation on integrating XML-based security into Web services*. In GCC Conference & Exhibition, 2009 5th IEEE. IEEE,1–5.

[5] Feifan Liu, Jinying Chen, Abhyuday Jagannatha, Hong Yu, *"Learning for Biomedical Information Extraction: Methodological Review of Recent Advances"*, arXiv: Computation and Language Journal, pp: 1-10, 2016

[6] S Pouriyeh, M Doroodchi, and M Rezaeinejad. 2010. *Secure Mobile Approaches Using Web Services*. In Conference: Proceedings of the 2010 International Conference on Semantic Web & Web Services, July, pp:12-15, Las Vegas, Nevada, USA, 201

[7] Xu Y, Hong K, Tsujii J, et al. *Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries*. J Am Med Inform Association; pp: 824–832, 2012.

[8] de Bruijn B, Cherry C, Kiritchenko S, et al. *Machine-learned solutions for three stages of clinical information extraction: the state of the art* , 2010. J Am Med Inform Assoc; pp: 557–562, 2013.

[9] Patra R, Saha SK. *A kernel-based approach for biomedical named entity recognition*. ScientificWorldJournal 2013; 2013:950796.

[10] Zhang Y, Lin H, Yang Z, et al. *Hash sub-graph pair wise-kernel for protein-protein interaction extraction*. IEEE/ACM Trans Computer Bioinform, pp:1190–1202, 2012.

[11] Thomas P, Neves M, Solt I, et al. *Relation extraction for drug-drug interactions using ensemble learning. 1st Challenge task on Drug-Drug Interaction Extraction*, pp: 11–18, 2013.

[12] Yang Z, Tang N, Zhang X, et al. *Multiple kernel learning in protein-protein interaction extraction from biomedical literature*. Artif Intell Med; 51:163–173, 2014.
[13] Li L, Zhang P, Zheng T, et al. *Integrating semantic information into multiple kernels for protein-protein interaction extraction from biomedical literatures*, 2014.
[14] Poon H, Vanderwende L. *Joint inference for knowledge extraction from biomedical literature. Human Language Technologies*: *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*; 813–821, 2017
[15] Hassani, H, *"A review of data mining applications incrime. Statistical Anal. Data Mining Rserach"*, ASA Data Sci Journal, pp: 139 – 154, 2016.
[16] Hearst, M.A, *"Untangling text data mining"*, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp: 3 – 10, 2009.
[17] T. Joachims, *"Text Categorization with Support Vector Machines"*, Technical Report 23, Universitat Dortmund, LS VIII, 20017
[18] Anurag Sarkar, Saptarshi Chatterjee, Writayan Das, Debabrata Datta, *"Text Classification using Support Vector Machine"*, International Journal of Engineering Science Invention, ISSN (Online): pp: 2319 – 6734, Volume 4 Issue 11, November 2015.
[19] Fatimah Wulandini, Anto Satriyo Nugroho, *"Text Classification Using Support Vector Machine for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Diseases"*, International Conference on Rural Information and Communication Technology, 2009.