Document Clustering on Large-Scale Data Using Ultra-Scalable Spectral Clustering and Ensemble Clustering

[1] Sandhya R. Jadhav, [2] Prof S. S. Redekar
[1]*M. Tech. Student, Computer Science and Engineering, AMGOI, Maharashtra, India*
[2]*Asst. Professor, Computer Science and Engineering, AMGOI, Maharashtra, India*

***Abstract* - Now a days huge amount of information is available, but it is definitely an important task to find relevant information from these. There are many ways to find relevant information from huge amount of data. Automatic data clustering systems play important role in this task. These systems help to find relevant information and also help to organize according to degree of relevancy with the given query. The problem arises to organize data to classify which document are relevant and which are irrelevant. Here propose a two novel algorithms of data clustering using ultra-scalable spectral clustering (USPEC) and ultra-scalable ensemble clustering (U-SENC) apply on huge amount of data. By applying algorithms, we can classify the required data using these algorithms.**

***Index Terms* - Data clustering, Large-scale clustering, Spectral clustering, Ensemble clustering, Large-scale datasets.**

I.INTRODUCTION

Day by day mass of available information to us is increases. Though huge amount of information is available, but all is not relevant for us. For our use it works with most advantages, there we need the devices that can sort data, store data, and also investigate all the information which is accessible. It is not possible physically to read and arrange data physically. We must choose any alternative option for it. We can use automatic text categorization methods on retrieve relevant information.

When we are using existing upgrading, approaches give unsatisfactory predictive performances as well as poor scalability when handling complex database schemas and noisy or numeric values in real-world applications. However, “flattening” strategies tend to want considerable time and energy for the information transformation, end in losing the compact representations of the normalized databases and produce an especially large table with huge number of additional attributes and various NULL values (missing values). As a result, these difficulties have prevented a wider application of multi relational mining and post an urgent challenge to the info mining community. to handle the abovementioned problems, this text introduces a Descriptive clustering approach where neither “upgrading” nor “flattening” is required to bridge the gap between propositional learning algorithm sand relational.

In the approach of document clustering, Data analysis techniques, like clustering it will be wont to identify subsets of information instances with common characteristics. Users can explore the info by examining some instances in each group rather than instead of examining the instances of the whole data set. this enables users to focus efficiently on large relevant subsets Data sets, particularly for document collections. specifically, the descriptive grouping consists of automatic grouping sets of comparable instances in clusters and automatically generates an outline or a synthesis which will be interpreted by man for every group. the outline of every cluster allows a user determine there each of the group without having to look at its content for text documents, an outline suitable for every group may be a multi-word tag, an extracted title or an inventory of characteristic words.

Selection of features is an important pre-processing method used to rule out noisy features. The measurements of the features are reduced and data is much better understood and cluster results, efficiencies and performance are improved. It is widely used in fields such as the classification of text. It is thus used primarily to improve the efficiency and efficiency of clusters. Term frequency (TF), inverse document frequency (Tf · IDF) and their hybrids are the most commonly used function selection metrics Each document in the corpus consists of k characteristics with the highest selection of metric scales, according to the best methods of choosing, and some of the improvements are made in old methods. Documentation methods include binary (presence or absence of the document), TF (i.e. frequency of the document term), and TF.IDF. We are applying clustering algorithms in the final stage of the document clustering process, grouping the target documents on the basis of features selected into distinct clusters. Approaches for the document clustering are:

Data pre-processing: Data pre-processing is a method of data mining that involves changing raw data into a reasonable format. Real data in certain practices or drifts are regularly fragmented, conflicted or affected and are likely to contain numerous errors. The pre-processing of data is a shown strategy for solving these problems. Crude data for further preparation is provided by pre-processing of data. In fact, data is filthy and insufficient in relation to characteristic estimates, which are short on particular characteristics of the intrigue or which contain entire data. Data that contains mistakes or abnormalities are upsetting. They are contradictory because there is incoherence in codes or names. There are no quality data, so quality mining results are not available. The choice of quality must be based on data quality. The data centre needs predictable value data reconciliation

Stages of pre-processing in clustering: it's critical to underline that getting from an accumulation of documents to a clustering, is anything but a solitary activity, yet is more a procedure in numerous stages. These stages incorporate progressively conventional data recovery tasks, for instance, creeping, ordering, weighting, separating then forth. a little of those different procedures are key to the standard and execution of most clustering calculations, and it's during this way important to contemplate these stages together with a given clustering calculation to outfit its actual potential.

Pre-processing techniques, stop words removal: it's the initial stage of pre-processing that makes a pretty list of terms for the document. Word stops are the words that are sifted within the preparation of language information. The paper is examined to search out what percentage words are described.

Stemming: It is a process to reduce different words (common form) to their roots. Stemming is a strategy to find methods for search terms to improve the retrieval adequacy and to reduce the indexing of files. Stemming is typically achieved through expulsion from file terms of any add-ons and prefixes (appends) prior to the actual assignment of the term to the index. Since to means the structure of the word but is different, it is important to differentiate between each structure of the word and its structure. All sorts of stemming calculations were created to do this. For example: The word “like” has its forms like likes, likely, liking, liked.

II. LITERATURE REVIEW

Literature survey is that the most significant step in any quite research. Before start developing, we want to review the previous papers of our domain which we are working and on the premise of study we will predict or generate the downside and begin working with the reference of previous papers. during this section, we briefly review the related work on Data Clustering and their different techniques.

Dong Huang, Chang-Dong Wang, Jian-Sheng Wu, Jian-Huang Lai, Chee-KeongKwoh, “Ultra Scalable Spectral Clustering and Ensemble Clustering”: [1] Author introduces the information analysis techniques, like clustering it will be accustomed identify subsets of information instances with common characteristics.

L. He, N. Ray, Y. Guan, and H. Zhang: [2] Author propose an efficient spectral clustering method for large-scale data. the most idea in our method consists of employing random Fourier features to explicitly represent data in kernel space. The complexity of spectral clustering thus is shown not up to existing Nystrom approximation so large-scale data.

J. S. Wu, W. S. Zheng, J. H. Lai, and C. Y. Suen: [3] Author introduce a Euler clustering approach. Euler clustering employs Euler kernels so as to intrinsically map the input file onto a posh space of the identical dimension because the input or twice, so Euler clustering can get obviate kernel trick and doesn't have to depend upon any approximation or sampling on kernel function/matrix, whilst performing a more robust nonlinear clustering against noise and outliers. Moreover, since the initial Euler kernel cannot generate anon- negative similarity matrix and thus is inapplicable to spectral clustering, author introduce a positive Euler kernel, and more importantly we've proved when it can generate a non- negative similarity matrix. Author applies Euler kernel and also the proposed positive Euler kernel to kernel k-means and spectral clustering so on develop Eulerk mean sand Euler spectral clustering, respectively.

N. Iam-On, T. Boongoen, S. Garrett, and C. Price: [4] This paper presents a replacement link-based approach to bolster the standard matrix. this can be achieved using the similarity between clusters that are estimated from a link network model of the ensemble. It uses specifically new link based three algorithms which are proposed for the underlying similarity assessment. The last word clustering result's generated from the refined matrix using two different consensus functions of feature based and graph-based partitioning. This approach is that the first to cater to and explicitly employ the link between input partitions, which has not been emphasized by recent studies of matrix refinement. The link-based approach effectiveness is demonstrated over 10 datasets (synthetic and real) and three valuation measures. J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen: [5] during this paper, author provide a scientific study of K-means-based Consensus Clustering (KCC). Here specifically discussed a necessary and sufficient condition for utility functions which work for KCC.

H. Liu, J. Wu, T. Liu, D. Tao, and Y. Fu: [6] Author propose Spectral Ensemble Clustering (SEC) to leverage the benefits of co-association matrix in information integration but run more efficiently. We disclose the theoretical equivalence between SEC and weighted K-means clustering, which dramatically reduces the algorithmic complexity. We also derive the latent consensus function of SEC, which to our greatest knowledge is that the first to bridge co-association matrix-based methods to the methods with explicit global objective functions. Further, we prove in theory that SEC holds the robustness, generalizability and convergence properties. We finally extend SEC to fulfil the challenge arising from incomplete basic partitions, supported which a row-segmentation scheme for giant data clustering is proposed.

J.-T. Chien, describe the “Hierarchical theme and topic modelling,” [7] in this Taking into ac- count hierarchical data sets within the body of text, like words, phrases and documents, here structural learning is performed and also themes for sentences and words from a group of documents respectively. the connection between arguments and arguments in numerous data groupings is explored through an unsupervised procedure without limiting the number of clusters. A tree branching process is presented to draw the proportion so the topic for various phrases. They build a hierarchical the mean athematic model, which flexibly represents heterogeneous documents using non-parametric Bayesian parameters. The thematic phrases and also the thematic words are extracted. within the experiments, the proposed method is evaluated as effective for the development of a semantic tree structure for the corresponding sentence sand words. Thus, up priority of the utilization of the tree model

For the choice of expressive phrases for the summary of documents is illustrated. Bernardini, C. Carpineto, and M.D’ Amico, describe the “Full-subtopic retrieval with key phrase- based search results clustering,” in this Consider the matter of restoring multiple documents that are relevant to the individual sub-topics of a given Web query, called” full child retrieval”. to resolve this problem, they present a brand-new algorithm for grouping search results that generates clusters labelled with key phrases. The key phrases are extracted generalized suffix tree created by the search results and merge through a hierarchical agglomeration procedure improved grouping. They also introduce a brand-new measure to judge the performance of full recovery sub-themes, namely” hunt for secondary arguments length under the sufficiency of k documents”. they need used a test collection specifically designed to judge the recovery of the sub-themes, they need found that our algorithm has passed both other clustering algorithms of existing research results as a way of redirecting search results underline the variety of results (at least for k 1, that's once they have an interest in recovering quite one relevant document by sub-theme).

R. Xu and D. Wunsch, describe the “Survey of clustering algorithms,” in that Data analysis plays an in dispensable role in understanding the various phenomena. Conglomerate analysis, primitive exploration with little or no previous knowledge, consists of research developed in a wide variety of communities. Diversity, on the one hand, provides us with many tools. On the other hand, the profusion of options causes confusion. They have examined the grouping algorithms for the data sets that appear in statistics, computer science and machine learning and they illustrate their applications in some reference datasets, the problem of street vendors and bio informatics, and a new field that attracts intense efforts. Various closely related topics, proximity measurement and cluster validation are also discussed

III. SYSTEM ARCHITECTURE

In Proposed System training is creation of train data set using which clustering of unknown data in predefined categories is done. Here a learning system is created using advanced clustering algorithms. It is an advanced learning where unlabelled data is classified using labelled data. Training data is always a labelled dataset based on its features.

Project had considered no of scientific papers form different publication of different do- mains for creating training dataset. These papers are input for creating training dataset. This input is first pre-processed and most informative features are extracted using TF/IDF algorithm and word embedding sematic score algorithm. Ten different domains from market are identified and then extracted feature and have to put to corresponding domain where each domain is considered as one class that which is used for labelling test dataset in testing part and features are considered as nodes. Once training part is completed, all features of respective domains are get updated in corresponding tables in database.

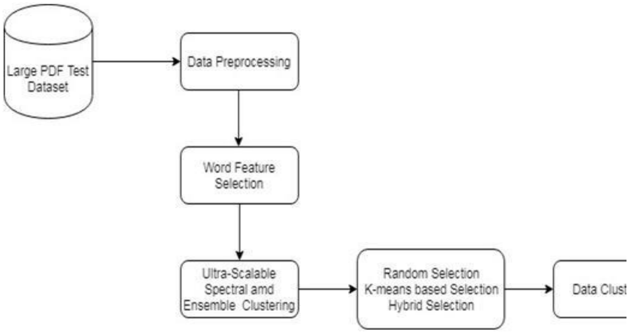

Large PDF Test Dataset
Data Preprocessing
Word Feature Selection
Ultra-Scalable Spectral amd Ensemble Clustering
Random Selection K-means based Selection Hybrid Selection
Data Clust

Figure 1. System Architecture

System Block Diagram

In our proposed system, Training set and testing set are main two modules present. Here from collected dataset major dataset is first pass to training set to train the model. After that according to requirement the dataset is passes to testing dataset.

Selected text is passes for pre-processing phase. In this phase our text goes through different four steps like stop word removal, stemming, TF-IDF and score calculator. After applying pre-processing on selected data, it moves toward final processing.

In final phase feature extraction is applied on data then data is ready for clustering. Classification is done here according to domain.

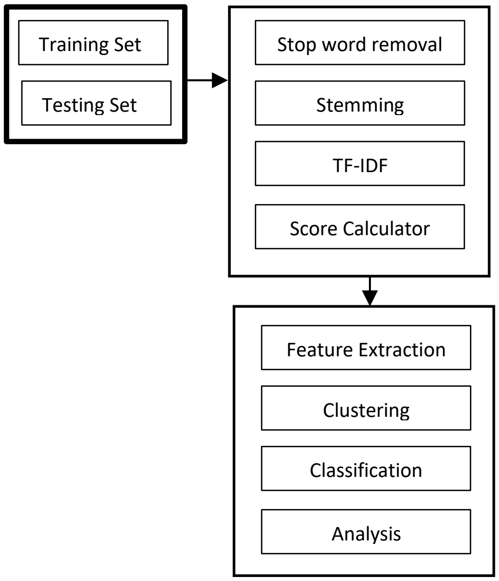
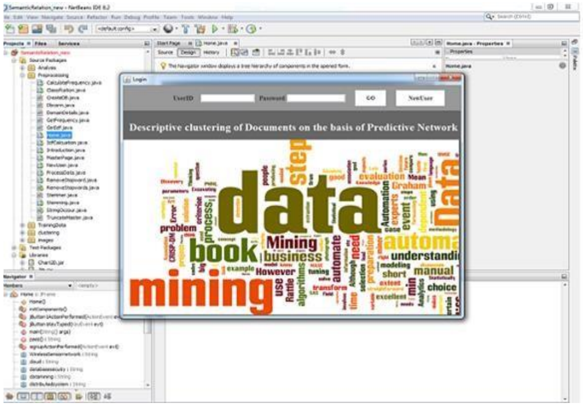


Figure 2. System Block Diagram

IV. RESULTS

Admin Home Page

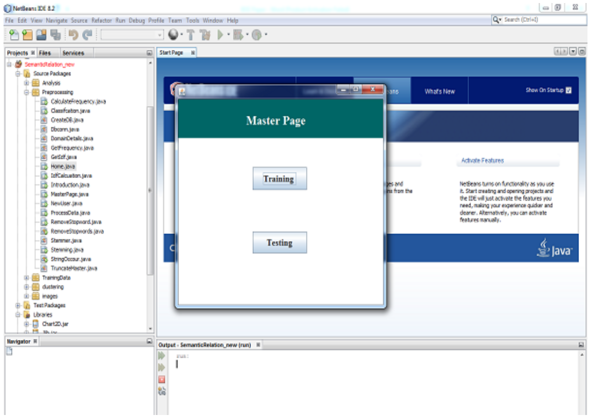
Here user must be complete registration phase for authenticated login process. After completion of this phase user can move toward further process the dataset for clustering. New user must register and old user can directly login by valid username and password.


UserID
Password
GO
NewUser

Descriptive clustering of Documents on the basis of Predictive Network

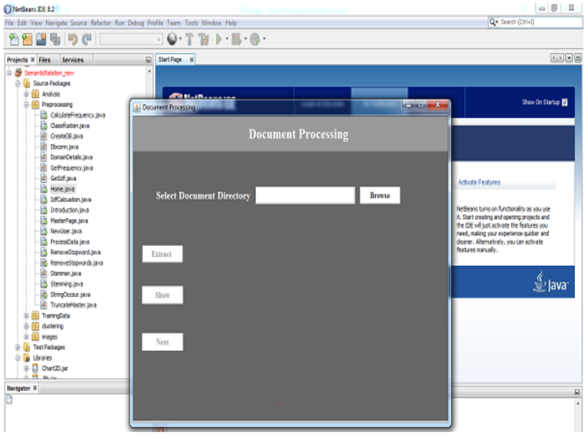
Module Page

After completing valid login user can choose training or testing dataset according user's choice for clustering.



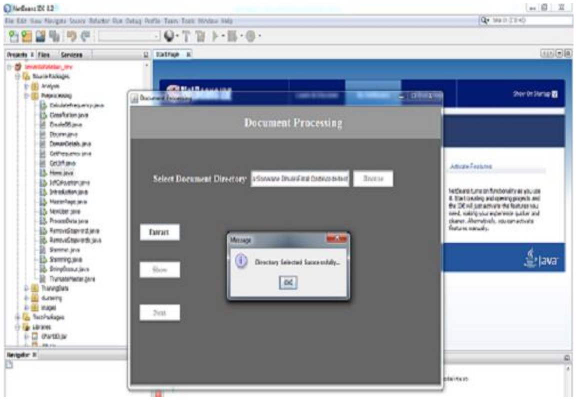
Select Input Folder Page

Once user enter from master page, user must select document dictionary to extract required data.

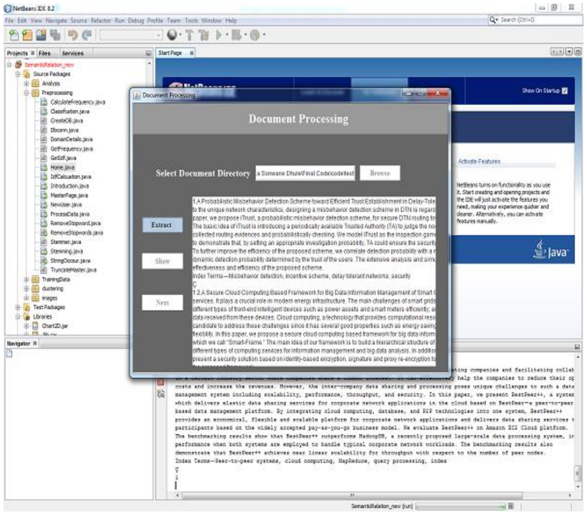


Extract Directory Page

Data from selected directory page is extracted here. Then the data is moving the further Process of pre-processing.



Directory Selected Successfully...



a Sonwane Dhule\Final Code\codetest

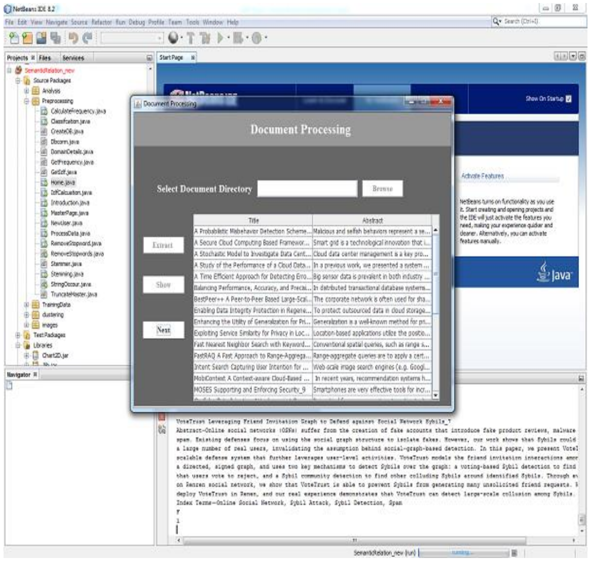
1.A Probabilistic Misbehavior Detection Scheme toward Efficient Trust Establishment in Delay-Tole
to the unique network characteristics, designing a misbehavior detection scheme in DTN is regard
paper, we propose iTrust, a probabilistic misbehavior detection scheme, for secure DTN routing to
The basic idea of iTrust is introducing a periodically available Trusted Authority (TA) to judge the no
collected routing evidences and probabilistically checking. We model iTrust as the inspection gam
to demonstrate that, by setting an appropriate investigation probability, TA could ensure the securi
To further improve the efficiency of the proposed scheme, we correlate detection probability with a
dynamic detection probability determined by the trust of the users. The extensive analysis and sim
effectiveness and efficiency of the proposed scheme.
Index Terms—Misbehavior detection, incentive scheme, delay tolerant networks, security

1 2.A Secure Cloud Computing Based Framework for Big Data Information Management of Smart
services. It plays a crucial role in modern energy infrastructure. The main challenges of smart grids
different types of front-end intelligent devices such as power assets and smart meters efficiently, a
data received from these devices. Cloud computing, a technology that provides computational reso
candidate to address these challenges since it has several good properties such as energy saving
flexibility. In this paper, we propose a secure cloud computing based framework for big data inform
which we call "Smart-Frame." The main idea of our framework is to build a hierarchical structure of
different types of computing services for information management and big data analysis. In additio
present a security solution based on identity-based encryption, signature and proxy re-encryption t

```
...ting companies and facilitating collab
...y help the companies to reduce their op
costs and increase the revenues. However, the inter-company data sharing and processing poses unique challenges to such a data
management system including scalability, performance, throughput, and security. In this paper, we present BestPeer++, a system
which delivers elastic data sharing services for corporate network applications in the cloud based on BestPeer—a peer-to-peer
based data management platform. By integrating cloud computing, database, and P2P technologies into one system, BestPeer++
provides an economical, flexible and scalable platform for corporate network applications and delivers data sharing services t
participants based on the widely accepted pay-as-you-go business model. We evaluate BestPeer++ on Amazon EC2 Cloud platform.
The benchmarking results show that BestPeer++ outperforms HadoopDB, a recently proposed large-scale data processing system, in
performance when both systems are employed to handle typical corporate network workloads. The benchmarking results also
demonstrate that BestPeer++ achieves near linear scalability for throughput with respect to the number of peer nodes.
Index Terms—Peer-to-peer systems, cloud computing, MapReduce, query processing, index
Ç
1
```

Show Paper Title and Abstract Page

Here selected paper titles and abstracts are display.

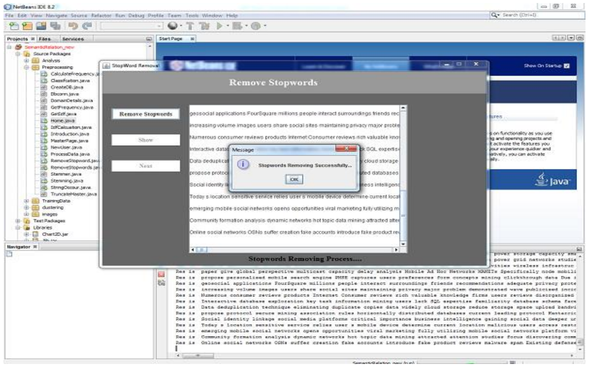


Title	Abstract
A Probabilistic Misbehavior Detection Scheme...	Malicious and selfish behaviors represent a se...
A Secure Cloud Computing Based Framewor...	Smart grid is a technological innovation that i...
A Stochastic Model to Investigate Data Cent...	Cloud data center management is a key pro...
A Study of the Performance of a Cloud Data...	In a previous work, we presented a system ...
A Time Efficient Approach for Detecting Erro...	Big sensor data is prevalent in both industry ...
Balancing Performance, Accuracy, and Precisi...	In distributed transactional database systems...
BestPeer++ A Peer-to-Peer Based Large-Scal...	The corporate network is often used for sha...
Enabling Data Integrity Protection in Regene...	To protect outsourced data in cloud storage...
Enhancing the Utility of Generalization for Pri...	Generalization is a well-known method for pr...
Exploiting Service Similarity for Privacy in Loc...	Location-based applications utilize the positio...
Fast Nearest Neighbor Search with Keyword...	Conventional spatial queries, such as range s...
FastRAQ A Fast Approach to Range-Aggrega...	Range-aggregate queries are to apply a cert...
Intent Search Capturing User Intention for ...	Web-scale image search engines (e.g. Googl...
MobiContext A Context-aware Cloud-Based ...	In recent years, recommendation systems h...
MOSES Supporting and Enforcing Security_9	Smartphones are very effective tools for incr...

```
VoteTrust Leveraging Friend Invitation Graph to Defend against Social Network Sybils_7
Abstract—Online social networks (OSNs) suffer from the creation of fake accounts that introduce fake product reviews, malware
spam. Existing defenses focus on using the social graph structure to isolate fakes. However, our work shows that Sybils could
a large number of real users, invalidating the assumption behind social-graph-based detection. In this paper, we present Vote
scalable defense system that further leverages user-level activities. VoteTrust models the friend invitation interactions amo
a directed, signed graph, and uses two key mechanisms to detect Sybils over the graph: a voting-based Sybil detection to find
that users vote to reject, and a Sybil community detection to find other colluding Sybils around identified Sybils. Through e
on Renren social network, we show that VoteTrust is able to prevent Sybils from generating many unsolicited friend requests.
deploy VoteTrust in Renen, and our real experience demonstrates that VoteTrust can detect large-scale collusion among Sybils.
Index Terms—Online Social Network, Sybil Attack, Sybil Detection, Spam
Y
1
```

Remove Stop word Page:

Stop words are removed here from extracted document. This is first step of pre-processing apply on data.

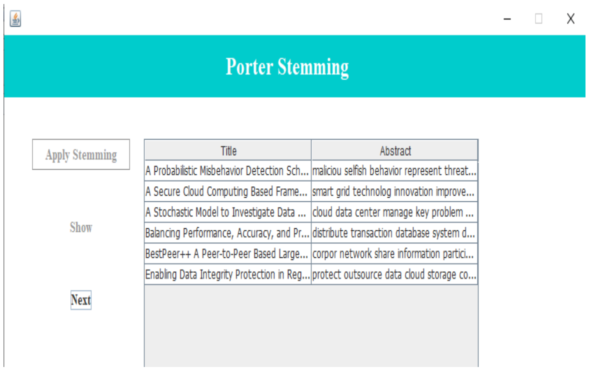


Stopwords Removing Successfully...

geosocial applications FourSquare millions people interact surroundings friends rec
increasing volume images users share social sites maintaining privacy major proble
Numerous consumer reviews products Internet Consumer reviews rich valuable kno
Interactive data...
Data deduplica... cloud storage
propose protoc... uted databases
Social identity li... ness intelligenc
Today s location sensitive service relies user s mobile device determine current locat
emerging mobile social networks opens opportunities viral marketing fully utilizing m
Community formation analysis dynamic networks hot topic data mining attracted atte
Online social networks OSNs suffer creation fake accounts introduce fake product re

```
Res is  paper give global perspective multicast capacity delay analysis Mobile Ad Hoc Networks MANETs Specifically node mobili
Res is  propose personalized mobile search engine PMSE captures users preferences form concepts mining clickthrough data Due
Res is  geosocial applications FourSquare millions people interact surroundings friends recommendations adequate privacy prote
Res is  increasing volume images users share social sites maintaining privacy major problem demonstrated wave publicized inci
Res is  Numerous consumer reviews products Internet Consumer reviews rich valuable knowledge firms users reviews disorganized
Res is  Interactive database exploration key task information mining users lack SQL expertise familiarity database schema fac
Res is  Data deduplication technique eliminating duplicate copies data widely cloud storage reduce storage space upload bandw
Res is  propose protocol secure mining association rules horizontally distributed databases current leading protocol Kantarci
Res is  Social identity linkage social media platforms critical importance business intelligence gaining social data deeper u
Res is  Today s location sensitive service relies user s mobile device determine current location malicious users access rest
Res is  emerging mobile social networks opens opportunities viral marketing fully utilizing mobile social networks platform v
Res is  Community formation analysis dynamic networks hot topic data mining attracted attention studies focus discovering com
Res is  Online social networks OSNs suffer creation fake accounts introduce fake product reviews malware spam Existing defens
```

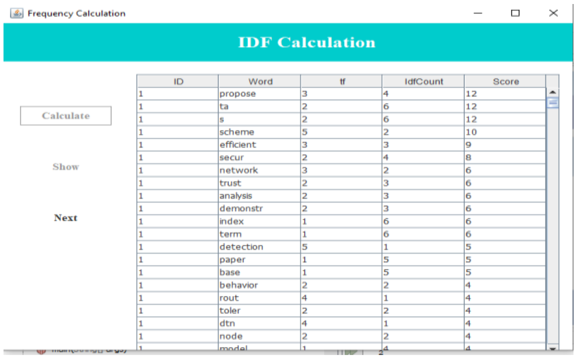
After applying stemming



Title	Abstract
A Probabilistic Misbehavior Detection Sch...	maliciou selfish behavior represent threat...
A Secure Cloud Computing Based Frame...	smart grid technolog innovation improve...
A Stochastic Model to Investigate Data ...	cloud data center manage key problem ...
Balancing Performance, Accuracy, and Pr...	distribute transaction database system d...
BestPeer++ A Peer-to-Peer Based Large...	corpor network share information partici...
Enabling Data Integrity Protection in Reg...	protect outsource data cloud storage co...

Final Score calculation

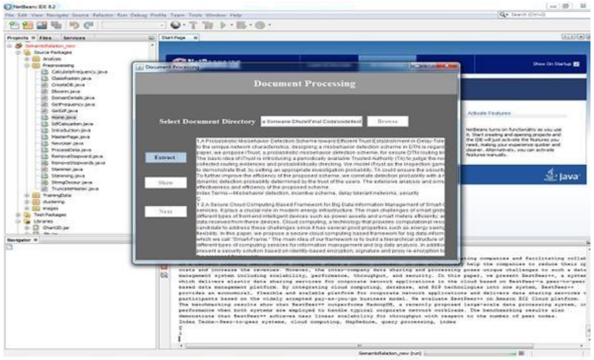
After calculating tf and idf the score will be calculated with formula tf * idf



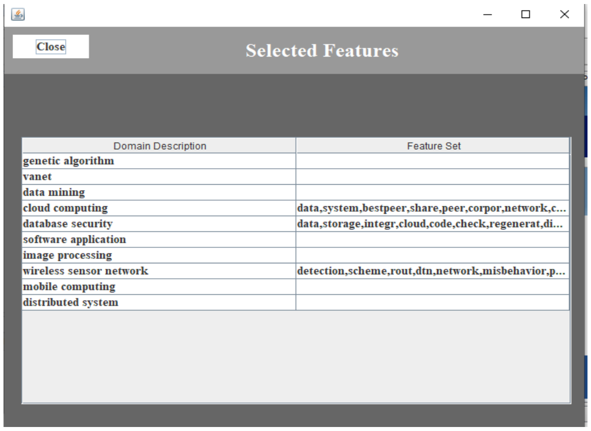
ID	Word	tf	IdfCount	Score
1	propose	3	4	12
1	ta	2	6	12
1	s	2	6	12
1	scheme	5	2	10
1	efficient	3	3	9
1	secur	2	4	8
1	network	3	2	6
1	trust	2	3	6
1	analysis	2	3	6
1	demonstr	2	3	6
1	index	1	6	6
1	term	1	6	6
1	detection	5	1	5
1	paper	1	5	5
1	base	1	5	5
1	behavior	2	2	4
1	rout	4	1	4
1	toler	2	2	4
1	dtn	4	1	4
1	node	2	2	4
1	model	1	4	4

Document Extracted:

clustering parameters the number of clusters and the number of functions per Clusters: they are chosen from the model selection. With the result solution, each group is described by a minimum subset of features necessary to predict if an instance belongs to the cluster our hypothesis is that even a user will be able to predict membership in the group of documents using the descriptive features selected by the algorithm. Given Some relevant requirements, a user can quickly identify clusters that probably contain relevant documents



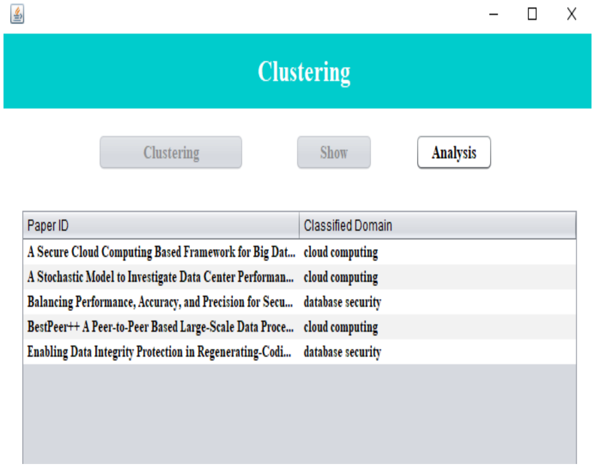
Features selected:



Domain Description	Feature Set
genetic algorithm	
vanet	
data mining	
cloud computing	data,system,bestpeer,share,peer,corpor,network,c...
database security	data,storage,integr,cloud,code,check,regenerat,di...
software application	
image processing	
wireless sensor network	detection,scheme,rout,dtn,network,misbehavior,p...
mobile computing	
distributed system	

Clustering Done:

Paper are domain wise separated. And shown in table.



Paper ID	Classified Domain
A Secure Cloud Computing Based Framework for Big Dat...	cloud computing
A Stochastic Model to Investigate Data Center Performan...	cloud computing
Balancing Performance, Accuracy, and Precision for Secu...	database security
BestPeer++ A Peer-to-Peer Based Large-Scale Data Proce...	cloud computing
Enabling Data Integrity Protection in Regenerating-Codi...	database security

V. PREPROCESSING ALGORITHMS

1. Stop word Removal-This technique remove stop words like is, are, they, but etc.
2. Tokenization-This technique remove Special character and images.
3. Stemming – remove suffix and prefix and Find Original word for e.g.-

1. Played – play

2.Clustering – cluster

TFIDF Algorithm

The tf–idf score for term I^i_j is calculated using the term frequency and the document frequency. Term frequency (tf) and inverse document frequency (idf) are the foundations of the most popular term weighting scheme in IR. The tf–idf score of I^i_j, tf—idf(I^i_j), is computed as:

$$\text{tf}-\text{idf}\left(l_i^j\right) \quad \log\left(tf\left(l_i^j, d_j\right)+1\right) \times \log\left(|D|/1+df\left(l_i^j, D\right)\right)$$

In experimental results, we evaluate the proposed system on student conference papers datasets this available on internet. We compare the accuracy of existing system results with proposed system.

The experimental result evaluation, we have notation as follows:

TP: True positive (correctly predicted number of instance)

FP: False positive (incorrectly predicted number of instance),

TN: True negative (correctly predicted the number of instances as not required)

FN false negative (incorrectly predicted the number of instances as not required),

On the basis of this parameter, we can calculate four measurements

Accuracy = TP+TN÷TP+FP+TN+FN

Precision = TP ÷TP+FP

Recall= TP÷TP+FN

F1-Measure = 2×Precision×Recall ÷Precision+ Recall.

VI. ACKNOWLEDGEMENT

The authors would like to thank all the authors mentioned in the references as well as all other authors for their earlier work on this topic.

VII. CONCLUSION

Proposed descriptive Clustering as two coupled predictions activity choose a grouping that is predictive of features and prediction of the cluster assignment of a subset of features. Use predictive performance as a goal criterion, descriptive clustering parameters the number of clusters and the number of functions per Clusters: they are chosen from the model selection. With the result solution, each group is described by a minimum subset of features necessary to predict if an instance belongs to the cluster our hypothesis is that even a user will be able to predict membership in the group of documents using the descriptive features selected by the algorithm. Given Some relevant requirements, a user can quickly identify clusters that probably contain relevant documents

REFERENCES

[1] Dong Huang, Chang-Dong Wang, Jian-Sheng Wu, Jian-Huang Lai, Chee-Keong Kwoh, “Ultra Scalable Spectral Clustering and Ensemble Clustering”, IEEE TRANSACTIONS ON KNOWLEDG AND DATA ENGINEERING VOL.12 NO.01 MAY 2019.

[2] L. He, N. Ray, Y. Guan, and H. Zhang, “Fast largescale spectral clustering via explicit feature mapping,” IEEE Trans. Cybernetics, in press, 2018.

[3] J. S. Wu, W. S. Zheng, J. H. Lai, and C. Y. Suen, “Euler clustering on large-scale dataset,” IEEE Trans. Big Data, in press, 2018.

[4] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, “A link-based approach to the cluster ensemble problem,” IEEE Trans. PAMI, vol. 33, no. 12, pp. 2396–2409, 2011.

[5] J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, “Kmeans-based consensus clustering: A unified view,” IEEE Trans. KDE, vol. 27, no. 1, pp. 155–169, 2015.

[6] H. Liu, J. Wu, T. Liu, D. Tao, and Y. Fu, “Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence,” IEEE Trans. KDE, vol. 29, no. 5, pp. 1129–1143, 2017.

[7] J.-T. Chien, “Hierarchical theme and topic modeling,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 3, pp. 565– 578, 2016.

[8] Bernardini, C. Carpineto, and M. D’Amico, “Fullsubtopic retrieval with keyphrase- based search results clustering,” in IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intelli- gent Agent Technol., 2009, pp. 206–213.

[9] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, “Self-organization of a massive document collection,” IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 574–585, 2000.

[10] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, “A hierarchi- cal monothetic document clustering algorithm for summarization and browsing search results,” in Proc. Int. Conf. World Wide Web, 2004, pp. 658–665.

[11] R. Xu and D. Wunsch, “Survey of clustering algorithms,” IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645– 678, 2005.