IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 06, Issue 12 (Dec. 2016), ||V2|| PP 36-39

Efficient Clustering Of Text Documents Using Term Based Clustering

N.Deepika[1], K.Poornimathi[2], J.Anitha[3], A.Irumporai[4]
[1,2,3,4] Assistant Professor,Department Of Information Technology, Rajalakshmi Engineering College, Chennai.

Abstract: Text Mining Is The Practice Of Automatic Analysis Of One Document Or A Collection Of Documents And The Extraction Of Non-Trivial Information From The Documents. The Unstructured Text Data Is Processed Using Text Mining To Extract A Structured Data. Now-A-Days, There Exists A Problem In Segregating The Files Into Various Categories, Since It Is An Unstructured Data. Traditional Data Mining Provides Many Algorithms For Mining Useful Patters From Text Documents. But These Derived Patterns Are Not Fruitfully Used For Any Knowledge Discovery Process Especially In The Text Mining Domain. This Paper Presents An Effective Pattern Discovery Technique That Clusters The Documents Based On The Term Frequency Using Frequent Term-Based Clustering. This Will Aid In Improving The Effectiveness Of Updating And Applying The Discovered Patterns In Order To Find The Relevant Information And Knowledge. This Proposed Method Has Also Achieved A Good Performance Level When Compared To Other Existing Techniques.

Key Words: *Text Mining, Text Categorization, Term Based Clustering, Term Frequency*.

I. 1 INTRODUCTION

Text Mining Is An Escalating Field That Attempts To Extract Meaningful Information From A Huge Text Data. But, In Modern World, Text Is The Most Common Source For The Formal Exchange Of Information. The Field Of Text Mining Deals With Texts Whose Function Is The Communication Of Factual Information Or Opinions, And The Motivation For Trying To Extract Information From Such Text Automatically. Text Mining Is A Variation On A Field Called Data Mining, Which Tries To Find Interesting Patterns From Large Databases. High-Quality Information Is Typically Derived From Dividing The Patterns And Trends Over Methods Such As Statistical Pattern Learning. Text Mining Usually Involves The Process Of Converting The Input Text Into A Structured Format, Deriving Patterns With The Structured Data, And Involves Evaluation And Interpretation Of The Output At The End Of The Process. 'High Quality' In Text Mining Usually Refers To Some Combination Of Relevance, Novelty, And Interestingness.

Earlier In This Decade, A Large Number Of Data Mining Techniques Have Been Presented In Order To Perform Different Knowledge Discovery Tasks. These Techniques Include Association Rule Mining, Sequential Pattern Mining, Frequent Term Set Mining, Closed Pattern Mining, And Maximum Pattern Mining. Most Of These Techniques Are Proposed For Proposed To Develop Efficient Mining Algorithms To Generate Particular Patterns. With A Large Number Of Patterns Generated By Using Data Mining Approaches, How To Effectively Use And Update These Patterns Is Still An Open Research Issue. In This Paper, We Focus In A Developing A Knowledge Discovery Model To Effectively Use And Update The Discovered Patterns And Apply It To The Field Of Text Mining.

II. RELATED WORKS

Divya Nasa[1] Made A Survey Of Different Text Mining Techniques That Are Used In Various Fields Of Data Analytics. The Overall Framework Of Text Mining Is Explained And It Was Applied On A Sample Text Data To Obtain The Clustered Data. Rainee Kaptain And Jamp Kamps [2] Developed A System To Summarize The Search Engine Result Pages (Serp's). The Study Was Made With Three Word Cloud Generation Methods Such As Full-Text, Query Biased And Anchor Text Based Clouds And Finally Evaluated Them Using A Used Study. They Obtained A 70% Correspondence Of The Subtopic Judgement And 60% Of Relevance Assessment Judgement. The Accuracy Of The Results Was Still Low Even Though Three Word Cloud Generation Methods Were Followed In This System.Weiwei Et Al.[3] Used A Visualization Model That Couples A Trend Chart With A Word Cloud For The Purpose Of Temporal Content Evaluation In A Set Of Documents. This Involves Both The Generation And Coupling Of Word Cloud And Trend Chart, But This Result Was Not Effectively Used To Label The Data Under Categories Over The Use Of Adaptive Force Directed Model Tend To Change The Structure And Nature Of The Data After Formation Of Word Cloud.Ning Zhong Et Al.[4] Analysed Various Text Mining Techniques. The Work Was Based On Developing An Effective Pattern Discovery Technique To Overcome The Low-Frequency And Misinterpretation Problems In Text

Mining. This Method Failed To Focus On The Grouping And Labelling Of Text Documents Based Upon The Frequency Of Word In The Document.Vishwanath Bijalwan Et Al.[5] Used A Knn Based Machine Learning Algorithm To Categorize The Documents Which Will Finally Retrieve The Most Relevant Documents. Use Of Traditional Data Mining Techniques Will Consume A Lot Of Time To Categorize The Data Based Upon The Content In The Document And Finally Retrieving The Documents. This Method Is Time Consuming And Inefficient. Ranveer Kaur And Shruti Aggarwal[6] Explained The Various Text Mining Techniques And Also The Entire Process Involved In Text Mining. They Have Also Made A Detailed Survey On The Various Applications Of Text Mining In Various Field Of Development.

III. STAGES OF TEXT MINING

Text Mining Involves The Process Of Extracting High Quality Information From The Source By Using Various Data Mining And Text Mining Techniques. The Process Of Text Mining Involves Three Main Stages.

3.1 Information Retrieval: The First Stage Of Text Mining Is To Retrieve Information. This Requires The Use Of A Search Engine To Identify The Collection Of Words From The Texts That Are Already Processed Or It Might Require Pre-Processing Of Physical Texts. This Collection Of Words Will Need To Be Brought Together In A Useful Format.

3.2 Data Mining: It Is The Process Of Identifying Patterns In Large Sets Of Data, To Find That New Knowledge.

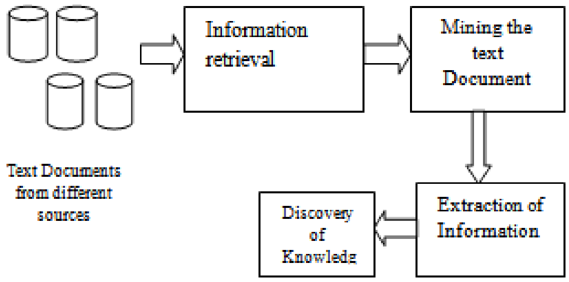

Fig 1: Text Mining Process

3.3 Information Extractions: In This Stage,

Information Extraction System Gathers Information From The Collection Of Words And Produces A Structured Representation Of Relevant Information I.E. The Relation Among The Clusters Or A Knowledge Base.

IV. TASKS OF TEXT MINING ALGORITHMS

4.1 **Text Clustering**: Clustering Is A Technique Used To Group Similar Documents But It Differs From Categorization In Than Documents Are Clusters On The Fly Instead Of Through The Use Of Pre-Defined Topics. (E.G. Self-Organizing Maps). In Our System We Follow The Frequent Term-Based Clustering Approach For Text Classification Provides A Way To Reduce The Large Dimensionality Of The Document Vector Space. The Proposed System Deals With The Term Set I.E. The Frequent Term Set Of The Document. In This Method Only The Low Dimensional Frequent Term Set Is Used As A Cluster Member. This Cluster Consists Of A Set Of Documents That Contains All The Terms Of A Frequent Term Set. In Our Technique We Propose To Use The Mutual Overlap Of The Frequent Term Sets In Accordance With Their Sets Of Supporting Documents To Determine A Cluster. The Reason Behind This Approach Is That A Minimal Overlap Of The Clusters Will Result In A Small Classification Error, When The Clustering Is Later Used For Classifying New Documents.For Each Global Frequent Term Set, An Initial Cluster Is Constructed To Include All The Documents Containing This Term Set. Initially The Clusters Are Overlapping Because One Document Contains Multiple Global Frequent Term Sets. This Method Uses This Global Frequent Term Set As The Cluster Label To Identify The Cluster. For Each Document, The Initial Cluster Is Identified And The Document Is Assigned To The Best Matching Initial Cluster.

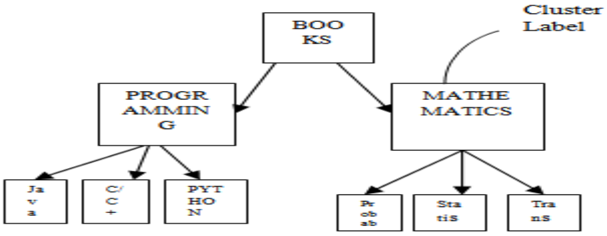

Fig 2: Clustered Text Documents Based On Term Frequency

Fig 2 Represents A Cluster Tree Where Each Cluster Other Than The Root Node Has Exactly One Parent. The Name Of A Parent Cluster Is More General Than The Name Of A Child Cluster And They Are "Similar" To A Certain Degree. Each Cluster Uses A Global Frequent K-Term Set As Its Cluster Label. A Cluster With A K-Term Set Cluster Label Appears At Level K In The Tree. The Cluster Tree Is Built Bottom Up By Choosing The "Best" Parent At Level K- 1 For Each Cluster At Level K. The Parent's Cluster Label Must Be A Subset Of The Child's Cluster Label. By Treating All Documents In The Child Cluster As A Single Document, The Criterion For Selecting The Best Parent Is Similar To The One For Choosing The Best Cluster For A Document. Example: Cluster {Java, C/C++, Python} Has A Global Frequent 3-Termset Label. Its Potential Parents Are {Programming} And {Books}.

4.2 Pruning Cluster Tree: The Cluster Tree Can Be Wide And Deep, Which Is Not Suitable For Searching. The Main Aim Of Tree Pruning Is To Efficiently Remove The Most Specific Clusters Based On The Concept Of Inter-Cluster Similarity. The Focus Is, If Two Sibling Clusters Are Very Similar, They Should Be Merged Into One Cluster. If A Child Cluster Is Very Similar To Its Parent (High Inter-Cluster Similarity), Then The Child Cluster Should Be Replaced With Its Parent Cluster. The Parent Cluster Will Then Also Include All Documents Of The Child Cluster.

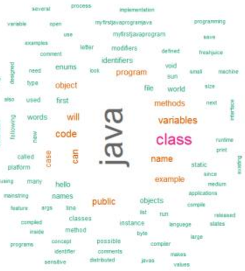

Fig 3: Word Cloud

The Clusters Obtained Were Used For Generating The Word Cloud In Order To Find The Global Frequent Term. Fig 3 Represents The Word Cloud That Shows "Java" As The Global Frequent Term.

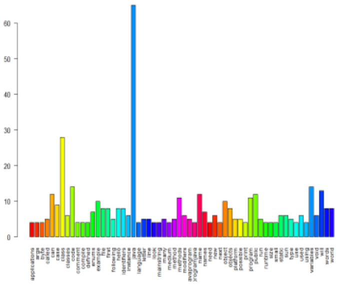

Fig 4: Word Frequency Graph

The Graph Depicts The Terms In The Document With Respect To The Frequency Of Occurrence.

4.3 Evaluation Of Frequent Term-Based Clustering: The Ftc Algorithm Was Experimentally Evaluated And Compared To Many Existing Document Clustering Methods. Ftc Uses Only The Global Frequent Terms In Document Vectors, Which Hugely Reduces The Dimensionality Of The Document Set. Research Works Show That Clustering Done With Reduced Number Of Dimensions Is Significantly More Efficient And Scalable. Ftc Can Cluster 100 Thousand Documents Within Several Minutes While Hierarchical Ftc And Upgma Cannot

Even Produce A Clustering Solution. The Obtained Cluster Tree Thus Provides A Logical Arrangement Of Clusters That Facilitates Searching Documents. Each Cluster Is Attached With A Cluster Label That Briefs The Documents In The Cluster. Different From Other Clustering Methods, No Separate Post-Processing Is Required For Generating These Meaningful Cluster Briefings.

4.4 Information Retrieval: Based On The Clusters Formed, The User Query Is Compared So That The Document With The Highest Number Of Global Frequent Terms Is Retrieved First And Is Provided To The User. .

4.5 Information Extraction: Text Documents Contain Information That Is Unstructured. As Documents Are Actually A Bag Of Words, They Can Be Represented By Vector Model Which Then Can Be Used As An Input To Techniques Such As Clustering Or Classifications. In Information Extraction, The Documents Are First Converted Into The Structured Format On Which Data Mining Techniques Are Applied To Extract Knowledge Or Interesting Patterns.

V. CONCLUSION

Most Of The Pattern Discovery Models For The Clustering Text Documents Uses Traditional Data Mining Techniques. This Proves To Be Inefficient Because, The Discovered Pattern Lacks Specificity Which Leads To Difficulty In Retrieving The Document. The Proposed System Uses Frequent Term-Based Clustering Approach That Effectively Clusters The Documents According To Frequent Occurrence Of Terms. This Method Generates A Clear Word Cloud That Can Be Used To Segregate The Documents, Thus Enabling Easy Retrieval Of Text Document Based On The User Query.

REFERENCES

[1] Divya Nasa – "Text Mining Techniques- A Survey "-International Journal Of Advanced Research In Computer Science And Software Engineering Research Paper Volume 2, Issue 4, April 2012

[2] Rianne Kaptein , Jaap Kamps- "Word Clouds Of Multiple Search Results"- Springer Lncs 6653,Pp 78-93, 2011.

[3] Weiweicui , Yingcaiwu- "Context Preserving Dynamic Word Cloud Visualization"- Ieee Pacific Visualisation Symposium 2010 2 - 5 March, Taipei, Taiwan

[4] Ning Zhong, Yuefeng Li, And Sheng-Tang Wu -"Effective Pattern Discovery For Text Mining "-Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012

[5] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari And Jordan Pascual- "Knn Based Machine Learning Approach For Text And Document Mining"- International Journal Of Database Theory And Application Vol.7, No.1 (2014), Pp.61-70

[6] Ranveer Kaur, Shruti Aggarwal-" Techniques For Mining Text Documents "- International Journal Of Computer Applications (0975 – 8887) Volume 66– No.18, March 2013