HIERARCHICAL DOCUMENT ORGANIZATION AND RETRIEVAL BASED ON THEMES FOR NEWS TRACKS

S. M. Arnica Sowmi
Assistant Professor
Department of Computer Science and Engineering
Alpha College of Engineering and Technology
Pondicherry
arni.sindhu@pec.edu

D. Dinesh Babu
Assistant Professor
Department of Computer Science and Engineering
Alpha College of Engineering and Technology
Pondicherry
dineshbabu89@pec.edu

***Abstract*—Organizing text documents is an important task and there are also numbers of strategies available in it. A good document clustering approach can assist computers in organizing the document corpus automatically into a meaningful cluster hierarchy for efficient browsing and navigation, which is very valuable for overcoming the deficiencies of traditional information retrieval methods. By clustering the text documents, the documents sharing the same topic are grouped together. Unlike document classification, no labelled documents are provided in clustering. Hence clustering is also known as unsupervised learning. In case of term based data retrieval, time consumption problem prevails. This is because as for each term, the data set's has to be retrieved. Hence we are going for taxonomy based data retrieval. This paper presents the taxonomical approach of clustering data set in a dynamic environment. It is a difficult task to cluster data in a dynamic environment. But this can be made easily by using RSS feeds.**

Keywords-ANITA approach; CHRONICLE construction; RSS feeds; Taxonomical clustering.

I. INTRODUCTION

Clustering text documents in a static environment can be done easily and the problem arises when it needs to be done in the dynamic environment. It becomes a difficult task if the number of text documents gets increased. This can be easily done, if the clustering is performed using taxonomical construction.

In a taxonomy-based information organization, each category in the hierarchy can index text documents that are relevant to it, facilitating the user in the navigation and access to the available contents. For a document collection whose content changes over time, a given initial taxonomy may soon lose its effectiveness in guiding users to relevant documents. In such cases, we revise the existing taxonomy in the light of new data. In the existing system, for clustering process, they have used this taxonomical approach as, ANITA [1] clustering approach. They have clustered science group of data sets. Once the documents are clustered, then the data retrieval is done. For data retrieval, a new algorithm as verb-only algorithm [2] is proposed. Here the data retrieval is done according to the user query. The user query will be given based on the occupational activities as clustered datasets.

It is seen that, in both the existing systems, the clustering as well as the data retrieval is done for static group of datasets. Hence, to assist the user search query and clustering process in a dynamic environment, the proposed system is designed. In the proposed system, the news group contents are clustered by using the ANITA [1] clustering approach with the refined steps. For assisting the user search query in a dynamic environment, the CHRONICLE construction is been proposed, which is also otherwise called as, verb-noun algorithm.

II. RELATED WORKS

Taxonomy facilitates the formalized knowledge for the organization of data and define aggregations for the various concepts in the particular domain .Thus it makes the contents easy access to the user. Many authors tried in building taxonomical hierarchies for the text corpus. In particular, [3] showed the traditional clustering approaches in document clustering as k-means and hierarchical agglomerative clustering. This HAC algorithm

is based on finding the inter object distances and then builds a binary tree hierarchy This suit only for static group of data sets and seems to be time consuming one. Further many authors tried in building concept hierarchies without use of training data or standard clustering techniques. In [4], the concept hierarchy is been built to find the association type among the concepts. By using this association type, the concept hierarchy is constructed. This method could not satisfy large group of datasets, which is found to be a drawback in this method.

Another method as, Faceted search, navigation and browsing [5] is proposed by K.-P. Yee, K. Swearingen, K. Li, M.Hearst. In this paper, the authors proved it as a popular information filtering technique for accessing a data collection represented using a faceted classification. The faceted classification enables classifications to be ordered in multiple ways, rather than in a single, pre-determined, taxonomical order. Apart from this an innovative approach is proposed for exploring text collections using a novel keywords-by-concepts (KbC) graph [6]. This supports navigation using domain-specific concepts as well as keywords which characterizes the text corpus. In [7], authors Kunal Punera and suju rajan introduces a new approach for extracting the hierarchical structure automatically from the text corpus. This technique discovers relationships among documents that are not encoded in the class labels. The relationships are represented in the form of sub-trees and SVM classifiers are used for classifying those nodes in the trees.

III. PROPOSED SYSTEM

The proposed system aims at clustering news groups of web contents in a dynamic environment. The RSS (Really Simple Syndication) feeds of news group websites are identified and the news contents are extracted. Once the contents are extracted, then they are pre-processed and the ANITA taxonomical approach [1] is implemented. This results in the taxonomical method of clustering news group contents. The RSS feeds provide the recent updated information's of any website. Here the RSS feeds of two news group websites [14], [15] are considered and their web contents are clustered. The overall architecture of the proposed system is shown below.

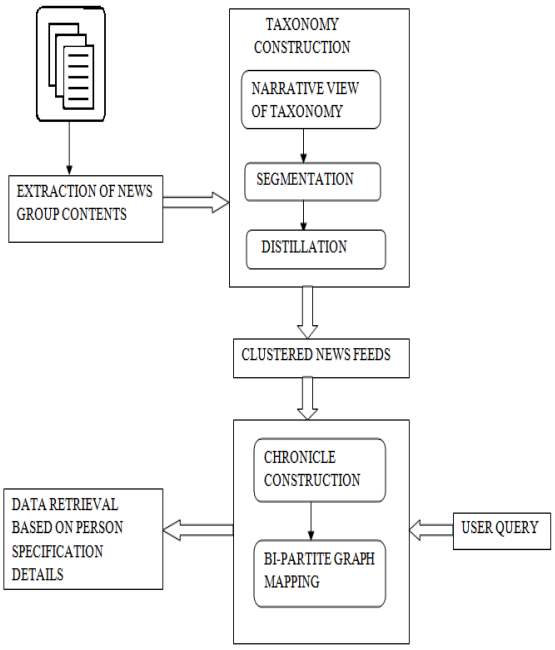

Figure1. PROPOSED SYSTEM ARCHITECTURE

The above architecture figure 1 clearly shows the entire process that is carried out during clustering process. There are three modules are found here. The first module includes the extraction of news group contents from the specified RSS feeds. Then the second module includes clustering those news contents using ANITA [10] clustering approach. The third module is the final step of the clustering process, where data retrieval is done. Here the data retrieval is done for person specification dataset.

- *Extraction of news feeds*

The clustering process begins with the extraction of news group contents. In order to extract the news group content dynamically, the rss feeds of the particular website should be known. In our clustering process, the news

feeds from two websites [14], [15] are used. The two news group websites considered for clustering process are as follows:

- http://rss.nytimes.com/services/xml/rss/nyt/World.xml
- http://feeds.feedburner.com/cnet/tcoc.xml

Once the contents are extracted, then they are organized by using ANITA clustering approach.

- *ANITA Clustering Approach*

The clustering process begins with the separation of title and description from the extracted news group contents. Then the ANITA taxonomical approach is implemented. Here, the ANITA clustering process [10], involves construction of taxonomical clusters with the considered dataset. The ANITA clustering process is slightly refined here. The following modules of the ANITA clustering approach are

- Narrative view of a taxonomy
- Segmentation of the narrative
- Taxonomy distillation / reconstruction

- *Narrative View of a Taxonomy*

In this process, it involves two main steps as,

- Concept sentences
- Sentence ordering

a) Concept Sentences – are the vectors obtained by analyzing the structure of the given taxonomy and the related corpus of documents. i.e., they are associated to each concept as a coherent set of semantically related keywords, extracted from the associated text corpus.

b) Sentence ordering – involves the creation of the narrative by selecting a permutation which captures the structure of the taxonomy as well as the content of the considered corpus. This can be based on three ancestor descendant ordering constraints, as

- **Pre – order constraint**
- **Post – order constraint**
- **Parenthetical constraint**

Pre – order constraint – here the root node is considered as the most general concept and the leaf node as the related terms. The sentences associated to the nodes of the taxonomy are read in pre–order.

Post – order constraint – generates the narrative in which the different concepts are presented bottom-up.

Parenthetical constraint – here the ancestor is repeated twice in the narrative. That is, each parent node is visited twice, representing both the general introduction and the conclusion to the argument that the children specialize.

- *Segmentation Of The Narrative*

In this step, we will be identifying the correlated segments or the concepts in the given corpus. The main idea is that, if two concepts are highly correlated then they need not be two separate nodes in the adapted taxonomy.

- *Taxonomy Distillation / Reconstruction*

In order to construct the adopted taxonomy from the partitions created in the previous step, we need to reassemble the partitions in the form of a tree structure. Thus for each partition we will be providing the label, which describes the concepts in that partition.

Once the clustering is done, then the documents need to be retrieved. The documents are retrieved based on the occupation related activities or the activities based on general population specified in the cluster. Clustering people into a smaller number of classes allows the grouping of practitioners of the occupations that share a considerable number of occupation related activities. Thus, analyzing descriptions of people belonging to various occupations, we can build a hierarchy of occupations. This entire process is implemented by using verb-noun algorithm or CHRONICLE construction.

IV. THE CHRONICLE CONSTRUCTION

The web contents which are extracted from the net are clustered in the form of taxonomical clusters. This chronicle construction is mainly done for the retrieval of person specific activities. The person specified in the particular news cluster is identified and their personal details as name, DOB, country, job specifications, and popularity were listed out. In the existing system, only the occupational activities are retrieved for the clustered documents. In this CHRONICLE construction, the internal mapping is done for extracting the person specification details. The steps for CHRONICLE construction are as follows:

The CHRONICLE construction steps easily retrieves the person specification details according to the user query provided. Once the query is given, the person name specified in the particular news cluster is mapped to the user query and then the person specification details are displayed. This CHRONICLE construction is also otherwise called as verb-noun algorithm. The algorithm is as follows:

1. Get the user query, Q is taken as input and then the algorithm is implemented.
2. Begin
3. Extraction of RSS feeds of respective news track links,
   http://rss.nytimes.com/services/xml/rss/nyt/World.xml
   http://feeds.feedburner.com/cnet/tcoc.xml
4. Displaying news as taxonomical clusters
5. Retrieval of person specification details, where PD includes, name, DOB, popularity, country, continent, job specification.
6. Constructing bipartite graph for PD retrieval, includes mapping of person details, where G= {N, V, E}, N= list of names, V = job specifications of individuals, E= arcs connecting the names and the activities.
7. Displaying PD details
8. End

The above algorithm clearly shows the actual process that is been carried out during the CHRONICLE construction process. Once this process is been implemented then the internal mapping is done for easily retrieving the user query informations. The retrieved data provides the personal specifications of any of the world politicians. The personal specifications are as follows:

- Job specifications
- Popularity
- Country
- Continent
- Hobbies

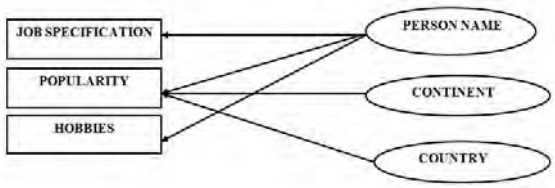

Figure 2. A set of Bi-partite graph containing person specification details

Here, for internal mapping of personal specification data's, the Bi-partite graph is constructed. As soon as the user query is given, the news feeds appears in the taxonomical clusters. If any politician name is depicted in the clustered news feeds, then the personal specification details as, job specification, DOB, Country, Continent, Hobbies details are displayed for that particular person.

A) EFFECTIVE MEASURES

In general, any document clustering can be evaluated by using the two major clustering evaluation techniques. These two evaluates the quality of the cluster as well as the coherence between the objects present under each cluster. The two evaluation techniques are as follows:

a) Purity - purity measure evaluates the coherence of a cluster, that is, the degree to which a cluster contains documents from a single category. Suppose for a given particular cluster C_i of size n_i, the purity of C_i is formally defined as :

$$\mathbf{P(C_i)} = \frac{1}{n_i \max\ h\,(n_i^h)} \quad \text{---------- (4.1)}$$

Where max h (n_i^h) is the number of documents that are from the dominant category in cluster C_i and n_i^h represents the number of documents from cluster C_i assigned to category h.

b) F measure – denotes the quality of the cluster. It is generally measured as the harmonic mean of recall (fraction of all relevant documents retrieved) and precision (fraction of retrieved documents that are relevant).Hence for any document, a cluster is treated as the result of the query and similarly each class as the desired set of documents for a query. Hence for for cluster j and class i,

$$\text{recall}(i,j) = n_{ij}/n_i \quad \text{---------- (4.2)}$$

$$\text{precision}(i,j) = n_{ij}/n_j \quad \text{---------- (4.3)}$$

$$F(i,j) = \frac{(2 * \text{recall}\ (i,j) * \text{precision}\ (i,j))}{((\text{precision}(i,j) + \text{recall}(i,j))} \quad \text{---------- (4.4)}$$

Where n_{ij} is the number of members of class i in cluster j, n_j is the number of members of cluster j and n_i is the number of members of class i. To evaluate the clustering results, precision, recall, and F-measure were calculated over pairs of points. For each pair of points that share at least one cluster in the overlapping clustering results, these measures try to estimate whether the prediction of this pair as being in the same cluster was correct with respect to the underlying true categories in the data.

c) Precision and recall - In general, Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. Precision is calculated as the fraction of pairs correctly put in the same cluster, and recall is the fraction of actual pairs that were identified.

In simple terms, high **recall** means that an algorithm returned most of the relevant results, while high **precision** means that an algorithm returned substantially more relevant results than irrelevant. The above table shows the f measure value calculated for the scientific literature documents as well as the news group contents. Here, each cluster consists of its own sub-clusters. The general formula to find out precision and recall values are as follows:

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

Suppose if the number of documents are 30, out of which 20 are relevant and the remaining 40 other documents are irrelevant, then the precision value will be, 20/30, which is 0.666. Similarly, the recall value will be, 20/60, which is 0.333.

B) Experimental Results

In the existing system, the science group of data sets are clustered by using the traditional clustering approaches as, hierarchical agglomerative clustering and k-means clustering algorithm. The obtained results is been compared with the proposed ANITA [1] clustering approach. The result shows that the ANITA clustering provides best results when compared to the other two clustering approaches.

TABLE I. COMPARATIVE STUDY OF BISECTING K-MEANS, HAC VS ANITA APPROACH FOR SCIENTIFIC LITERATURE AND NEWSGROUPS USING PURITY

SCIENTIFIC LITERATURE				**NEWS GROUPS**		
clusters	**Bisecting k-means**	**HAC**	**ANITA approach**	**Bisecting k-means**	**HAC**	**ANITA approach**
1	0.28	0.35	0.46	0.34	0.36	0.46
2	0.35	0.38	0.48	0.38	0.40	0.50
3	0.36	0.40	0.44	0.42	0.47	0.56
4	0.38	0.40	0.48	0.40	0.48	0.6
5	0.40	0.45	0.55	0.45	0.49	0.7
6	0.50	0.56	0.65	0.52	0.58	0.8

The above table clearly shows the purity value calculated for science group of documents. Here, the proposed ANITA approach is compared with the two clustering algorithms as, bisecting k-means and hierarchical agglomerative clustering algorithms. Nearly 6 scientific taxonomical clusters are taken as sample datasets and the purity values are calculated. Each scientific taxonomical cluster will consists of inner-sub clusters. The purity value is calculated by using the above formula. From the calculated purity value, it is seen that the ANITA approach gives the highest purity value when compared to other two algorithms.

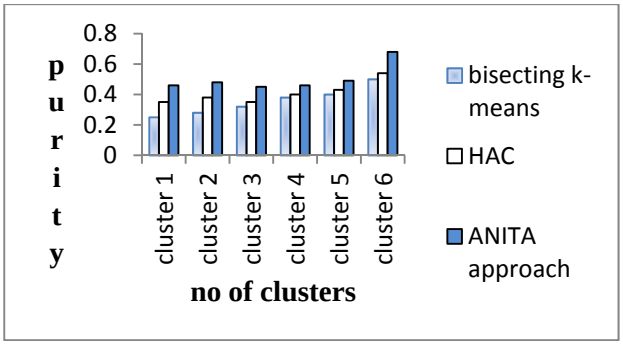

Figure 3. Comparative Study of Bisecting K-Means, HAC Vs ANITA Approach for Scientific Literature Using Purity

The above graph shows the comparative study of bisecting k-means and HAC with proposed ANITA approach for scientific literature using purity. Here, the scientific literature documents are taken as input and the clusters are formed. To the obtained clusters, the purity value is calculated. The values shows that the proposed ANITA approach gives the highest f-measure value when compared to bisecting k-means and HAC clustering algorithms.

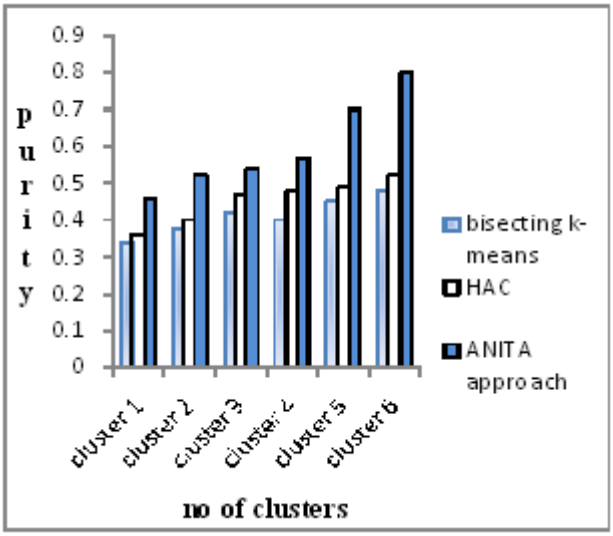

Figure 4. Comparative Study of Bisecting K-Means, HAC Vs ANITA Approach for News groups Using Purity

The above graph shows the comparative study of bisecting k-means and HAC with proposed ANITA approach for news group contents using f-measure. Here, the news group contents are taken as input and the clusters are formed. To the obtained clusters, the f-measure value is calculated. The values shows that the proposed ANITA approach gives the highest f-measure value when compared to bisecting k-means and HAC clustering algorithms.

TABLE II. COMPARATIVE STUDY OF BISECTING K-MEANS, HAC VS ANITA APPROACH FOR SCIENTIFIC LITERATURE USING F-MEASURE

SCIENTIFIC LITERATURE				NEWS GROUPS		
Clusters	**Bisecting k-means**	**HAC**	**ANITA approach**	**Bisecting k-means**	**HAC**	**ANITA approach**
1	0.28	0.35	0.46	0.34	0.36	0.46
2	0.35	0.38	0.48	0.38	0.40	0.50
3	0.36	0.40	0.44	0.42	0.47	0.56
4	0.38	0.40	0.48	0.40	0.48	0.6
5	0.40	0.45	0.55	0.45	0.49	0.7
6	0.50	0.56	0.65	0.52	0.58	0.8

The above table clearly shows the f-measure value calculated for science group of documents. Here, the proposed ANITA approach is compared with the two clustering algorithms as, bisecting k-means and hierarchical agglomerative clustering algorithms. Nearly 6 scientific taxonomical clusters are taken as sample datasets and the purity values are calculated. Each scientific taxonomical cluster will consists of inner-sub clusters. The f-measure value is calculated by using the above formula. From the calculated purity value, it is seen that the ANITA approach gives the highest f-measure value when compared to other two algorithms.

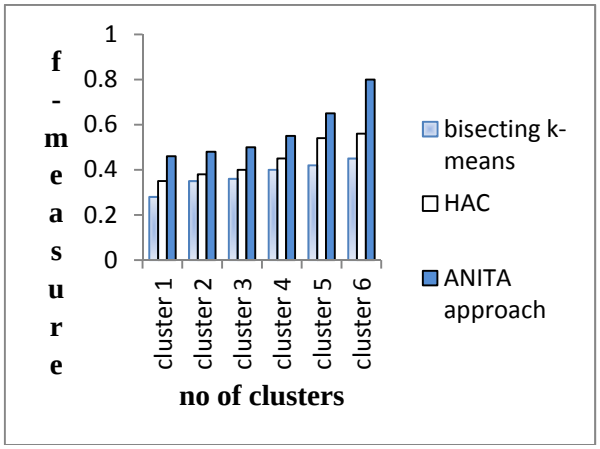

Figure 5. Comparative Study of Bisecting K-Means, HAC Vs ANITA Approach for Scientific Literature using f-measure

The below graph shows the comparative study of bisecting k-means and HAC with proposed ANITA approach for scientific literature using f-measure. Here, the scientific literature documents are taken as input and the clusters are formed. To the obtained clusters, the f-measure value is calculated. The values shows that the proposed ANITA.

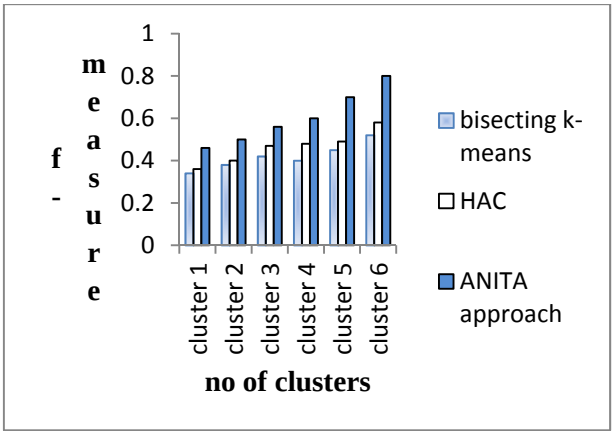

Figure 6. Comparative Study of Bisecting K-Means, HAC Vs ANITA Approach for Newsgroups using f-measure

V. CONCLUSION

As the number of documents in web gets increased, it is difficult task to perform clustering in a dynamic environment. The ANITA algorithm implemented in proposed system, helps in clustering dynamic group of documents in an efficient way. Here, for extracting the documents dynamically, the rss feeds of the particular websites are used. The rss feeds are defined to be really simple syndication, which provides the updated news as well as the simple description about the recent news.

After obtaining the taxonomical clusters, the data retrieval is done to assist the user query. This data retrieval is based on the retrieval of person specification details. In existing system, only the occupational activities are retrieved for the clustered data sets. In proposed system, the occupational as well as the person specification details are retrieved. This helps in the quick review about the particular person specified in the document cluster.

The proposed work can be further enhanced by obtaining the rss feeds for various domains as entertainment, science and even certain educational websites too. This helps in enhancing co-clustering of data.

REFERENCES

[1] M. Cataldi, K.S. Candan, M.L. Sapino, "ANITA: a narrative interpretation of taxonomies for their adaptation to text collections," Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM Conference on Information and Knowledge Management (CIKM), ACM, New York, USA, pp. 1781–1784, Oct 2010.

[2] Elena filatova, John Prager, "Occupation inference through detection and classification of biographical Activities," Department of Computer and Information Sciences, Fordham university, NY 10598,United States, Data and Knowledge Engineering, pp. 76-78, Aug 2012.

[3] B.S.Vamsi Krishna, P.Satheesh, Suneel Kumar R, "Comparative Study of K-means and Bisecting k-means Techniques in Wordnet Based Document Clustering," International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-6, August 2012.

[4] Phillipp Cimiano, Andreas Hotho, Steffen Stab, "Learning Concept Hierarchies From Text Corpora Using Text Corpora Using Formal Concept Analysis," Journal Of Artificial Intelligence Research (JAIR) 24, 305-339. Nov (2005).

[5] K.-P. Yee, K. Swearingen, K. Li, M. Hearst, "Faceted metadata for image search and navigation," Browsing Proceedings of CHI, ACM, pp. 401–408. Nov 2003.

[6] M. Cataldi, C. Schifanella, K.S. Candan, M.L. Sapino and L. Di Caro, Cosena, "A Context-Based Search And Navigation System," Proceedings of the International Conference on Management of Emergent Digital EcoSystems, MEDES'09, ACM, New York, NY, USA, 2009, pp. 218–225.

[7] K. Punera, S. Rajan, J. Ghosh, "Automatically learning document taxonomies for hierarchical classification," International World Wide Web Conference, ACM, pp. 1010–1011, Aug 2005.

[8] S. Dumais and H. Chen, " Hierarchical classification of web content," In SIGIR, pp 256–263, Nov 2000.

[9] Xujuan Zhou, Yuefeng Li, Peter Bruza, Yue Xu, Raymond Lau, "A Two-stage Information Filtering Based on Rough Decision Rule and Pattern Mining," Journal of Emerging Technologies In Web Intelligence (JETWI, ISSN 1798-0461), Nov 2006.

[10] Philipp Cimiano, Andreas Hotho, Steffen Staab, "Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis," Journal of Artificial Intelligence Research (JAIR) 24, 305–339. Nov (2005).

[11] L. Tang, H. Liu, J. Zhang, N. Agarwal, J.J. Salerno, "Topic taxonomy adaptation for group profiling,'' ACM Transactions on Knowledge Discovery from Data, pp.1-28, Aug (2008).

[12] http://nsdl.org,toorganizedigitalresources.

[13] http://www.dmoz.org/.

[14] http://rss.nytimes.com/services/xml/rss/nyt/World.xml.

[15] http://feeds.feedburner.com/cnet/tcoc.xml.

[16] http://en.wikipedia.org/wiki/Precision_and_recall#Precision .

AUTHORS PROFILE

ArnicaSowmi S.M. is an Assistant Professor at Alpha College of Engineering and technology, puducherry. She has completed M.Tech degree in 2013, at Pondicherry Engineering College, Puducherry. Her areas of interest are, Clustering, Cloud Computing, Sofware Engineering, Data Communication, Natural Language Processing. Currently she is doing research under Data Management in Pervasive Computing.

D. Dinesh Babu is an Assistant Professor at Alpha College of Engineering and technology, puducherry. He has completed his M.Tech degree in 2013, at Pondicherry Engineering College, Puducherry. His areas of interest are Network Security, Software architecture, Cloud Computing, Internet of Things, Clustering. Currently he is doing research under Data integrity in Cloud.