ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Modified Taxonomy Based Anita Approach for Dynamic Environment

Jayabharathy.J, Arnica Sowmi.M.S

Department of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry, India

ABSTRACT: Organizing text documents is an important task and there exists numbers of strategies in the recent research. 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 overcomes the drawbacks of traditional information retrieval methods. Through clustering, the documents sharing same topic are grouped together. Hence clustering is also known as unsupervised learning. In case of term based data retrieval, time consumption problem prevails. This paper concentrates on taxonomy based data retrieval. Taxonomies, which facilitates the organization of data in a given domain and make the contents easy access to the user This paper presents the taxonomical approach of clustering data set in a dynamic environment. The existing static ANITA approach has modified for dynamic environment and results are presented.

KEYWORDS: Taxonomical clustering, RSS feeds, CHRONICLE construction, ANITA approach

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 work

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

10.15680/ijircce.2015.0304152

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-septic 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 ALGORITHM

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 [1,9] 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.

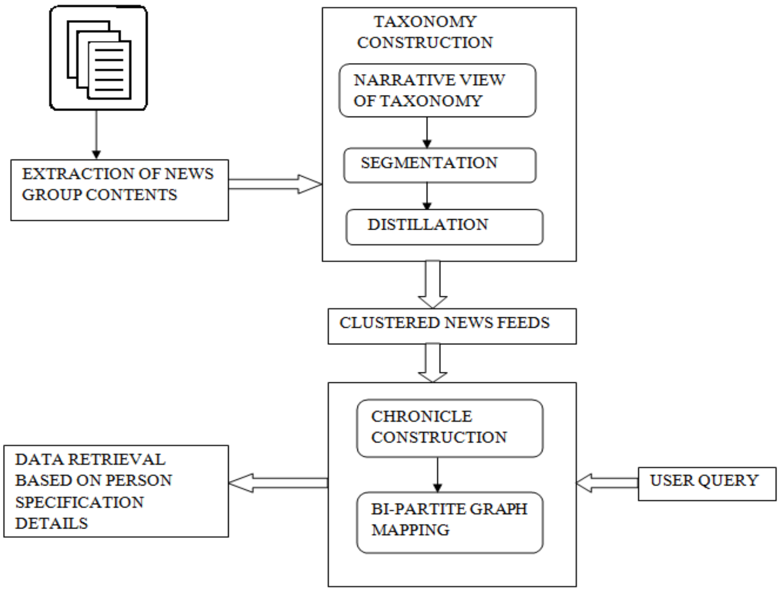

Fig. 1 Proposed System Architecture

The above architecture Fig. 1 clearly shows the entire process that is carried out during clustering process. Three main modules are of the proposed system are discussed. 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 [4] 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.

A. 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.

B. 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 are the modules of ANITA clustering approach:

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

i) Narrative View of a Taxonomy

In this process, it involves two main steps as,

- ✓ Concept sentences
- ✓ Sentence ordering

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.

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.

ii) 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.

iii) 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, analysing 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 information's. 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

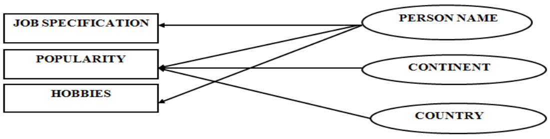

Fig. 2 A set of Bi-partite graph containing person specification details

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.

IV. EXPERIMENTAL RESULTS

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:

F-Measure and Purity are the metrics used to analyze the performance of the proposed term based and correlated concept based Clustering and Topic detection algorithms developed for static and dynamic corpora.

F-Measure combines the Precision and Recall from information retrieval process. Each cluster is treated as if it was the result of a query, and each class as if it was the desired set of documents, for a query. The recall and precision of that cluster for each given class are calculated. More specifically, F-Measure for cluster *j* and class *i* is computed as follows:

$$\text{Recall}(i,j)=\frac{n_{ij}}{n_i} \tag{4.1}$$

$$\text{Precision}(i,j)=\frac{n_{ij}}{n_j} \tag{4.2}$$

$$F(i,j)=\frac{2\times\text{Recall}(i,j)\times\text{Precision}(i,j)}{\text{Precision}(i,j)+\text{Recall}(i,j)} \tag{4.3}$$

where n_{ij} is the number of members of the 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*. For each class, only the cluster with highest F-Measure is selected. Finally, the overall F-Measure of a clustering solution is weighted by the size of each cluster:

$$F(S)=\frac{1}{n}\sum_{j=1}^{n}\frac{n_j}{\max(F(i,j))} \tag{4.4}$$

Purity measure evaluates the coherence of a cluster, that is, the degree to which a cluster contains documents from a single class. Given a particular cluster C_i of size n_i the purity of C_i is defined as:

$$P(C_i)=\frac{1}{n}\max(n_i^h) \tag{4.5}$$

where *max*(n_i^h) is the number of documents that are from the dominant class in cluster C_i and n_i^h represents the number of documents from cluster C_i assigned to class *h*. The overall purity of a clustering solution is:

$$\text{Purity}(S)=\frac{1}{n}\sum_{i=1}^{n}\max(n_i^h) \tag{4.6}$$

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.

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.

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.

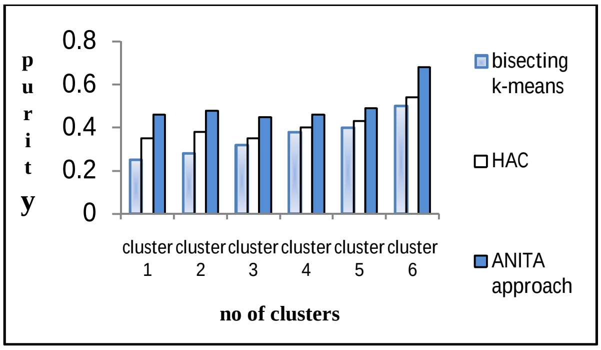

Fig. 4.2 Comparative Study of Bisecting K-Means, HAC Vs ANITA Approach for Scientific Literature Using Purity

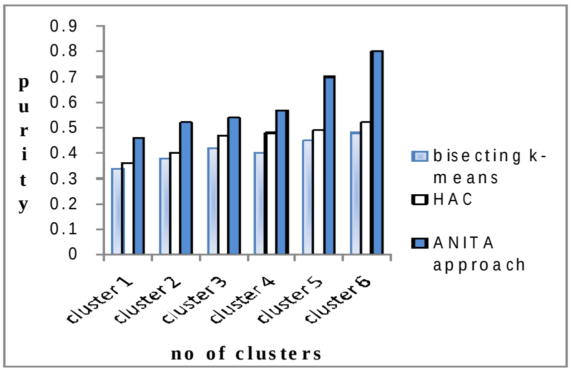

Fig. 4.3. Comparative Study of Bisecting K-Means, HAC Vs ANITA Approach for News groups Using Purity

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Fig 4.2 and 4.3 shows the comparative study of bisecting k-means and HAC with proposed ANITA approach for scientific literature and newsgroup dataset 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 Purity value when compared to bisecting k-means and HAC clustering algorithms.

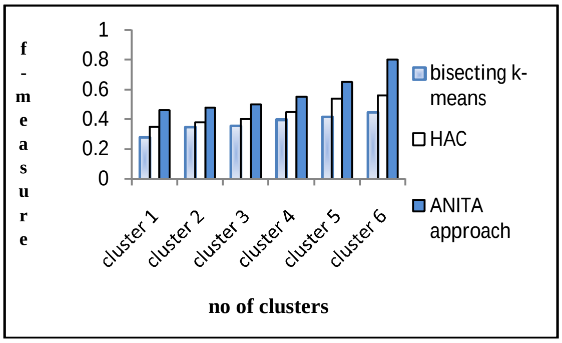

Fig. 4.4 Comparative Study of Bisecting K-Means, HAC Vs ANITA Approach for Scientific Literature using F-Measure

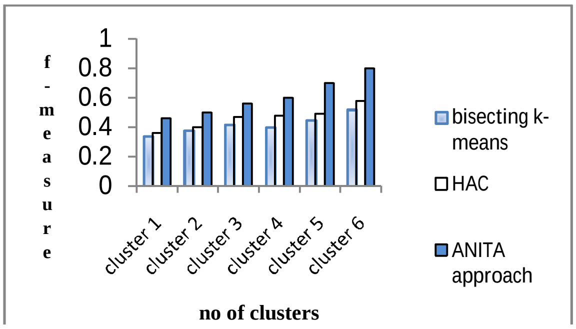

Fig. 4.5 Comparative Study of Bisecting K-Means, HAC Vs ANITA Approach for Newsgroups using F-Measure

Fig 4.4 and 4.5 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 show that the proposed ANITA approach gives the highest F-Measure when compared to bisecting k-means and HAC clustering algorithms.

From the graph, it is justified that the proposed algorithm gives better performance compared to the existing algorithms for scientific dataset. Also it is inferred that the existing ANITA approach for static environment has been modified and implemented for dynamic and results states that it works on par with the static bisecting and HAC algorithms.

V. Conclusion and Future Work

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, 2010.
2. Elena Filatova, John Prager, “Occupation inference through detection and classification of biographical Activities”, Data and Knowledge Engineering, Vol. 76-78,pp. 76-78, 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), Vol. 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) Vol. 24, pp.305-339, 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. 2003.
6. M. Cataldi, C. Schifanella, K.S. Candan, M.L. Sapino, 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, pp. 218–225, 2009.
7. K. Punera, S. Rajan, J. Ghosh, "Automatically learning document taxonomies for hierarchical classification", International World Wide Web Conference, ACM, pp. 1010–1011, 2005.
8. L. Tang, H. Liu, J. Zhang, N. Agarwal, J.J. Salerno, " Topic taxonomy adaptation for group profiling", ACM Transactions on Knowledge Discovery from Data", Vol. 4, pp.1-28, 2008.
9. http://nsdl.org,toorganizedigitalresources
10. http://www.dmoz.org/
11. http://rss.nytimes.com/services/xml/rss/nyt World.xml
12. http://feeds.feedburner.com/cnet/tcoc.xml
13. http://en.wikipedia.org/wiki/Precision_and_recall#Precision