Similarity of Articles using Hierarchical Clustering

Geetanjali, Shashank Sahu

Abstract— ***RSS technology is to find similarity in the articles to provide better services to user. The research is going on to find out semantic similarity in articles to reduce same type of articles read by user. Objective of RSS is to deliver content which is latest and consist of most relevant information to the user. Here, the research focus is to find out the suitable distance method that can be use to check similarity in the articles. Hierarchical clustering is one of the best methods to cluster the articles which are similar on some parameter various methods are used in HC (hierarchical clustering). Which method is best suitable to find the semantic similarity, similarity is the focus area of this paper. We have collected various articles from many news channel websites for a category (terrorist) by observing the articles. Thirty keywords are selected for the implementation for the proposed technique and comparison. We perform similarity checking on various numbers of articles like 18, 16, 12, 10, 9, and 6. After calculating the distance the cityblock distance method gives the best result. For this research work article from last one decade (2003-2013) has been selected.***

Keywords: Really Simple Syndication (RSS), Hierarchical Clustering.

I. INTRODUCTION

RSS means really simple syndication, Rich Site Summary and RDF (Resource Description Framework) Site Summary. In fact all the above techniques indicate the same techniques of syndication. Nowadays, RSS techniques are aggressively used in the field of online channels, wiki and blogs. By the using of RSS export techniques, users of the web site can subscribe to the news and not equal to quickly obtained information. To collect and classify the information provided by the RSS is also a very extremely interesting work. In the current "Information Age", internet affects our daily lives; we (people) use it as a tool to contact each other. Now a numerous organizations making use of it to supply information both genera as well as sensitive information like credit card information, financial information etc. to the users as well as in using mobile devices. RSS can also be used to distribute the latest information over the virtual space (Internet) [7], [17]. Hierarchical Clustering is also known as hierarchical cluster analysis or HCA. It is a method for analyzing clusters so that it builds a hierarchy of clusters. Hierarchical clustering are generally of two types:-

- **Bottom up Approach:** This is also known as "**Agglomerative**". In this, Each observation would start in their own cluster and then pair of cluster would be merged as one observation will move up in the hierarchy
- **Top down Approach:** This is also known as "**Divisive**" approach: In this, all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Usually, the merges and splits are determined in a greedy manner. The hierarchical clustering's results are generally represented in a dendrogram. The tree- based (dendrogram) structure of Hierarchal clustering shows in Figure 1.

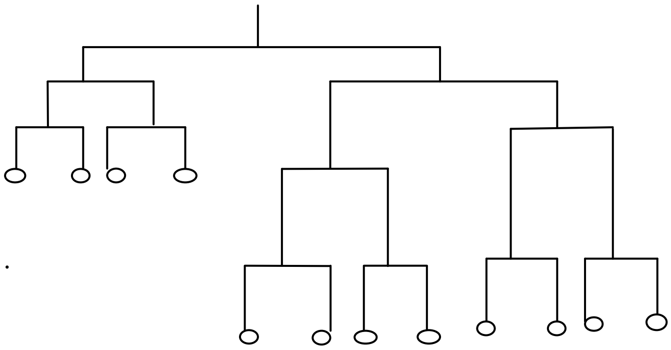

Figure 2.1 Dendrogram (Tree based structure)

RSS is a technique for delivering regularly changing web content. Many news related sites, weblogs and other online publishers syndicate their content as an RSS feed to whoever wants it. A user always requires up-to date content from online sources. The content is not always satisfactory for the user as it does not provide the specific information required by user. It is tedious task for common user to get the desirable information from vast Internet. To overcome this type of problem, the user needed RSS technology for ease of access. RSS gives the updated online information to the user. Problem in the research, RSS technology is to find similarity in the articles to provide better services to user. The research is going on to find out semantic similarity in articles to reduce same type of articles read by user. Objective of RSS is to deliver content which is latest and consist of most relevant information to the user. Here, the research focus is to finding suitable distance method that can be use to check similarity in the articles. Hierarchical clustering is one of the best methods to cluster the articles which are similar on some parameter various methods are used in HC (hierarchical clustering). Which method is best suitable to find the semantic similarity, similarity is the focus area of this paper. We have collected various articles from many news channel websites for a category (terrorist) by observing the articles. Thirty keywords are selected for the implementation for the proposed technique and comparison. We perform similarity checking on various numbers of articles like 18, 16, 12, 10, 9, and 6.

In the Matlab results are observed and show that cityblock method under the Hierarchical clustering (HC) is suitable method to finding similarity of articles for RSS feed. Cityblock method worked on frequency of words to calculate distance between articles. Cityblock is showing the results(c) in 0.9410, 0.9429, 0.9650, 0.9671, 0.9726 and 0.9760. c is named as cophenetic correlation coefficient. These values are more closer to 1 i.e cityblock gives the best results for finding similarity in articles than other distance method. Hierarchical clustering is applied on no. of articles and calculates the distance between the articles. In this first

Revised Version Manuscript Received on June 23, 2016.

Geetanjali, M.Tech, Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College Ghaziabad (U.P). India.

Shashank Sahu, Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College Ghaziabad (U.P). India.

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.*

collect the RSS data and then apply the hierarchical clustering. Here, five distance method take to calculate the distance like Euclidean, cityblock, jaccard, chebychev and minkowski.

The hierarchical clustering method will always combine two most similar groups into a single group and construct a new hierarchical structure. Apply hierarchical clustering. to matrix and get the Dendrogram. This Dendrogram hierarchy graph is constructed recursively. The contents that are close to each other are joined early and the nodes that have more difference are joined late. In this method a new cluster is created by combining two close clusters. This process continues until there is only one cluster. Usually we call it hierarchical cluster and it will give a dendrogram. The basic steps are:

Step 1: Computing all distances between all clusters;

Step 2:Finding out the closest clusters and forming a new cluster;

Step 3: Repeating step 1 and step 2 until only one cluster exists.

Distance is calculated using various methods like Euclidean Distance, Cityblock, Jaccard, Chebychev, Minkowski distance. There is much confusion which is the best for calculating distance between articles for RSS. Distance calculation is important element for RSS Feed technique. We have taken realistic articles from top News Channel like: Aajtak, ABP and so on for under the one category terrorist. The program in Java is made to count frequency of frequent words used in the articles that is needed in HC. A set of thirty words has been prepared by observing realistic articles. Various distance techniques has been applied on a matrix of thirty words using Matlab and results are noted down. Experiment in Matlab shows that cityblock distance in calculation in HC given the best result to find out similarity between articles for RSS Feed.

II. LITRATURE REVIEW

There are two techniques exist first one is Pull & second Push as shown in figure 2.1 as per Manfred Hauswirth. Mostly server client communication for distributed information system and browser uses request-reply model for communication. In Request- Reply Model client sends a request to the server to pull the information, while In push, client are registered with server for certain type of information and server broadcasts the information generated at regular interval to concerned clients [5].

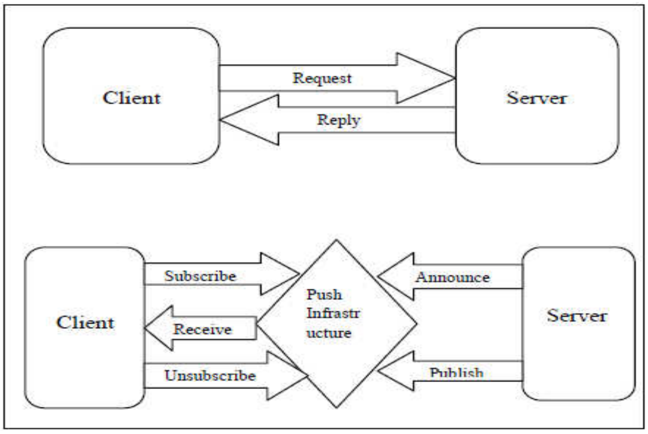

Figure 2.1 RSS Techniques [5]

Actually broadcaster applies some policies to filter the data to pass on the information data through channels and clients get the data as per their subscription [5]. As per Fekade Getahun Taddesse et al [6], by combining related or similar in some manner RSS news coming from one or different sources & providers, benefits and uses with different backgrounds. In this paper, author takes the problem and analyse the connectedness between RSS elements, on the basis of this analysis author come out with the most suitable relation between any two elements & provides the some predefined merging operators & adapted according to the human needs. According to Fekade Getahun et al [7], RSS query algebra improved in the direction of better news administration. Existing XML query algebras are misappropriated due to the following reasons: (1) RSS document is text leaded and content dependency on wording & authentication of author thus there is a need of operators named semantic-aware. (2) Dynamic & time retrieval of new items. (3) Overlapping of new & existing news through relationship identification. So the aim is to decipher the concerns/problem by delivering a dedicated RSS algebra based on semantic-aware operators which are capable of considering RSS characteristics. The application specific domain is provided by operators and could be varied depending on the p of the consumers. Facilitate the revision of set of queries and equivalence rules for over simplification & optimization. User could develop RSS query by using operator help of the EasyRSS-Manager [7]. By Wang Lei et al [8], a a new distributed algorithm of data compression from using hierarchical cluster model for sensor networks is proposed, the elementary ideas being, (a)Sensor network mapped into hierarchical cluster model. (b)Used of various wavelet transform models for data compression in inner and super clusters. (c)The comparative irregularity of sensor nodes set up & installed in super cluster.

By Peilin Shi [9], [10], a rough variable has behavior to group the gained fuzzy web access patterns. This characteristic of rough variable is used by rough k-means clustering algorithm. According the paper, measure the user interest by their visited web pages and time spend on each. This time is termed as fuzzy linguistic variables and web access pattern from web logs is changed to as fuzzy web access pattern. Fuzzy web access pattern is a fuzzy vector containing fuzzy linguistic variables or 0. Server service topic feed (right) as easily as normal RSS newsfeeds (left).

Markus M. Breunig et al. [11], proposed an intelligent compression technique and cluster only the compressed data records. Such compressed data records can be produced by the BIRCH algorithm. Typically they consist of the sufficient statistics of the form (*N*, *X*, *X*2) where *N* is the number of points, *X* is the (vector-) sum, and *X*2 is the square sum of the points. They can be used directly to speed up *k*-means type of clustering algorithms, but it is not obvious how to use them in a hierarchical clustering algorithm I-Ching Hsu [12], proposed Personalized Web Feeds Framework (PWFF) that is used to develop an Ontology-based Personalized Web Feed Platform (OPWFP). The rapid development of the social Web has resulted in diverse Web 2.0 applications for accessing various Web feeds such as Weblogs, news headlines, business products, real time information and Podcasts. Due to the masses of Web feeds available on the Internet, a major challenge is how to access them efficiently. Conventional methods of manually finding and matching keyword for Web feeds are time-consuming and inaccurate. It addresses this issue by defining a Personalized Web Feeds Framework (PWFF) for integrating ontology technologies into Web feeds and user profiles. The proposed OPWFP is provides customized Web feeds for personnel needs.

By Petros Belimpasakis et al. [13], the existing content sharing paradigms along with some advanced sharing use cases that are not feasible with the existing technologies. For satisfying these use cases, we propose a new system that allows content sharing in a totally user-centric manner, meaning that users can select the people they want to share their content with and just let the system handle all the lower level device, network bearer and content transfer technologies, which best fit each sharing occasion. The system feasibility is proved in two dimensions, firstly by a technical prototype implementation in a laboratory environment, and secondly via usability studies with non-expert users, for gathering their input and feedback on the interface and preferred interaction with such a system.

By Wen hu et al. [14], data clustering and analyzing techniques are studied by using hierarchical clustering method. A matrix of words is constructed with a randomly chosen RSS list. By collecting data from this list a matrix is built. In the matrix each row corresponds to a article and each column represents a word. Based on the matrix a hierarchical clustering algorithm is designed. In this algorithm the Pearson correlation coefficient is used to compute the distances among different contents. The dendrogram is used to describe the hierarchical relationship of contents and words. And the 2-D graph also is used to represent the dendrogram in another format.

III. PROPOSED METHODS

First we selected 18 articles and then we select one category which is “Terrorist”. In this we selected 18 articles that are selected from different news channel. We have developed a program to count the frequency of words. The program automatically counts frequency of words in the articles. We selected 30 words under one category “Terrorist”. After calculating frequency of words, distances between words are calculated using various method of hierarchical clustering for finding similarity between articles. We have selected one category “terrorist.

Steps of Word Count Program

In this, steps describe the process of the program to count the frequency of words & steps are following:

STEP 1: Read the name of the .txt files having the contents of articles stored in the folder named RSS.

STEP 2: The space separated list of the words to be search need to specify in the command line.

STEP 3: In the very first row we just write the list of searching words read from the command line.

STEP 4: Then pick the .txt file (Articles) one by one.

STEP 5: Start reading the file word by word using loop (for).

STEP 6: If read word is exists in the list of searching words list, we increase the counter for that word.

STEP 7: And continue the above step (step-6) till the end of file.

STEP 8: Print the counts for each searching word for that file (or article) in the next row.

STEP 9: Now pick another file and repeated the steps from step-5 to step-8 in the loop.

STEP 10: Do the same for all the .txt files in the RSS folder.

In this, first select a category of terrorist and take realistic articles from the news channels. And then make a matrix of words and articles; now apply HC for calculating distance by using different distance methods like ED, Cityblock, Jaccard, and so on. We get the dendrogram and the output c. c is the cophenetic correlation coefficient.

The closer the value of the cophenetic correlation coefficient is to 1, the more accurately the clustering solution reflects the data. User can use the cophenetic correlation coefficient to compare the results of clustering the same data set using different distance calculation methods or clustering algorithms to get the best results. Figure 4.4 shows the steps of implementation [15], [16].

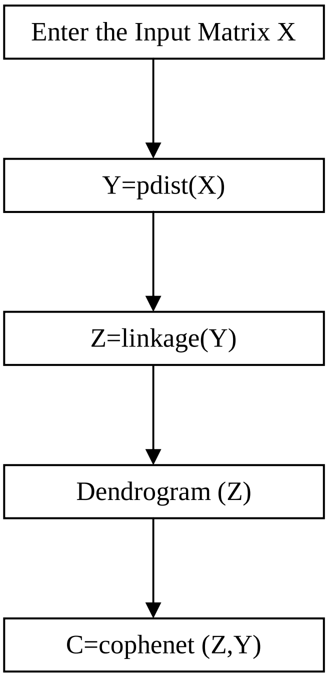

Figure 4.1 The Five Steps of Implementation

Figure 4.1 showing the sequence of implementation like: First we enter the matrix ‘x’ where ‘x’ is a matrix with two or more rows. The rows represent observations, the columns represent categories or dimensions and x is made up of 18/16/12/10/9/6 articles and 30 words. Then we calculate the pdist means pairwise distance between pairs of articles. Linkage’s function is; Agglomerative hierarchical cluster tree. After applying linkage function we calculate Cophenetic correlation coefficient (c). Finally similarity in articles is found by using HC for RSS feed.

IV. EXPERIMENT

In this section first we collect RSS content like 18, 16, 12, 10, 9 and 6 articles for implementation. And then formulation of matrix on the basis frequency of words. Application of various hierarchical distance methods to find out similarities among articles. Here we choose only one category that is “Terrorist” under thirty words. Analysis of result to find out efficiency of the applied method. Table 4.1 shows the best suitable method for finding similarity among articles is cityblock.

TABLE 4.1 Comparisons between Different no. of Articles by Using Different Distance Method

Methods/Articles	18 Articles	16 Articles	12 Articles	10 Articles	9 Articles	6 Articles
Euclidean	0.9225	0.9255	0.9498	0.9544	0.9655	0.9674
Cityblock	**0.9410**	**0.9429**	**0.9650**	**0.9671**	**0.9726**	**0.9760**
Jaccard	0.7974	0.8323	0.9225	0.9446	0.9513	0.8169
Chebychev	0.9027	0.9045	0.9315	0.9451	0.9458	0.9487
Minkowski	0.9255	0.9255	0.9498	0.9544	0.9655	0.9674

Table 4.1 shows that Cityblock distance method gives the best results in all cases (18, 16, 12, 10, 9 and 6), because it is closer to 1 means it gives the best result.

V. CONCLUSION

In this paper, we have shown that cityblock of Hierarchical clustering is suitable to find out similarity between articles. RSS works on similarity of content to deliver best possible content to users. We have experiment our proposed method on various articles like 18, 16, 12, 10, 9, and 6. The experiments have been performed on Matlab. It shows that cityblock is best method to find out the similarity between articles. We have started the experiment initially from six words and continuously increasing the words for better experiments. Finally we are able to finalize the 30 words under the category “Terrorist”. The reaserch work bounded with 30 words & one contextual category “terrorist”. The research work can be further extended for more meaningful words. It may include many categories like politics, customer survey, historical data, news on current technology etc. or future experiments.

REFERENCES

1. Den Ma, “Use of RSS feeds to push online content to users”, Published in Elsevier Journal of Decision Support Systems, Volume 54, Issue 1, December 2012, pp 740–749, doi: 10.1016/j.dss.2012.09.002.
2. Just van den Broecke, “Pushlets” Published in IEEE Transactions of Web Technology, Vol. 22, no. 5, may 2011.
3. Isabel delaTorre-Dıez, Saul Alvaro-Munoz, MiguelLo pez-Coronado, Joel Rodrigues, “Development and performance evaluation of a new RSS tool for a Web-based system: RSS_PROYECT”, Published in Elsevier Journal of Network and Computer Applications, Volume 36, Issue 1, January 2013, pp 255–261, doi: 10.1016/j.jnca.2012.06.004.
4. Lijing Zhang, “Research on Web-based Real-time Monitoring System on SVG and Comet”, Published in International Journal Vol.10, No.5, September 2012, pp 1142~1146, doi: 10.11591/telkomnika.v10i5.1347.
5. Manfred Hauswirth and Mehdi Jazayeri, “A Component and Communication Model for Push Systems”, Published in International Journal of Web Engineering and Technology, September 6-10, 1999**.**
6. Fekade Getahun Taddesse, Joe Tekli, Richard Chbeir, Marco Viviani & Kokou Yetongnon, “Semantic-based Merging of RSS Items”, Published in Springer Journal of World Wide Web March 2010, Volume 13, Issue 1-2, pp 169-207, doi: 10.1007/s11280-009-0074-4.
7. Fekade Getahun, Richard Chbeir, “RSS query algebra: Towards a better news management”, Published in Elsevier Journal of Information Sciences Volume 237, 10 July 2013, pp 313–342, doi: 10.1016/j.ins.2013.02.025.
8. Wang Lei, Wang Tongsen & Yang Ronghua, “ Data Compression Algorithm based on Hierarchical Cluster Model for Sensor Networks”, Published in International Conference of Future Generation Communication and Networking, 2008, Volume:2, pp 319-323, doi: 10.1109/FGCN.2008.96.
9. Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander, “Fast Hierarchical Clustering Based on Compressed Data and OPTICS”, Published in Proc. 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), Lyon, France**,** Volume 1910, 2000, pp 232-242 doi: 10.1007/3-540-45372-5_23.
10. I-Ching Hsu, “Personalized web feeds based on ontology technologies”, Published in Springer Journal of Information Systems Frontiers, July 2013, Volume 15, Issue 3, pp 465-479, doi: 10.1007/s10796-011-9337-6.
11. Petros Belimpasakis & Anne Saaranen, “Sharing with people: a system for user-centric content sharing”, Published in Springer Journal of Multimedia Systems, November 2010, Volume 16, Issue 6, pp 399-421, doi: 10.1007/s00530-010-0200-2.
12. Wen Hu & Qing he Pan, “Data clustering and analyzing techniques using hierarchical clustering method”, Published in Springer Multimed Tools Applications, July 2013, doi: 10.1007/s11042-013-1611-9.
13. John Garofalakis, Vassilios Stefanis, “Using RSS feeds for effective mobile web browsing”, published in Springer Journal of Universal Access in the Information Society, November 2007, Volume 6, Issue 3, pp 249-257, doi: 10.1007/s10209-007-0086-8.
14. Chen Wu, Elizabeth Chang, “Aligning with the Web: an atom-based architecture for Web services discovery”, published in Springer Journal 24 May 2007, doi: 10.1007/s11761-007-0008-x.
15. Saha, S, Sajjanhar, A., Shang Gao, Dew, R., Ying Zhao, "Delivering Categorized News Items Using RSS Feeds and Web Services", Published in International Conference of Computer and Information Technology (CIT), July 2010, pp 698-702, doi: 10.1109/CIT.2010.136.
16. Preechaveerakul, L, Kaewnopparat, W, "A Novel Approach: Secure Information Notifying System Using RSS Technology ", Published in International Conference of Future Networks, March 2009, pp 95-99, doi: 10.1109/ICFN.2009.35.
17. Segaran T, O’Brien MT (2007), “Programming Collective Intelligence, O’Reilly Media, Sebastopol”

Ms. Geetanjali Singh received her B.Tech degree in Computer Science & Engineering from ABES Engineering College, Ghaziabad (Gautam Budh Technical University, Lucknow), India, and pursuing M.Tech. in Computer Science & Engineering from Ajay Kumar Garg Engineering College Ghaziabad (Gautam Budh Technical University, Lucknow), India. Her main research interests are in Web technology, Client Server Computing, Algorithms.

Mr. Shashank Sahu is working as Associate Professor in Computer Science & Engineering Department at Ajay Kumar Garg Engg. College. Ghaziabad, India. He received his M.Tech degree in Computer Science & Engineering from G. B. Technical University, Lucknow, India. He is pursuing Ph.D. in Computer Science & Engineering from Sharda University, Greater Noida, India. He has 17 years of academic experience. His research areas are Software Engineering, Computer Architecture and Artificial Intelligence. He is the author of more than 8 publications in national/international conferences