Proceedings of the Pakistan Academy of Sciences:
A. Physical and Computational Sciences 53 (3): 281–291 (2016)

ISSN: 2518-4245 (print), 2518-4253 (online)

Pakistan Academy of Sciences

Research Article

Text Clusters Labeling using WordNet and Term Frequency-Inverse Document Frequency

Syed Muhammad Saqlain[1,*], Asif Nawaz[1], Imran Khan[1], Faiz Ali Shah[2], and Muhammad Usman Ashraf[1]

[1]International Islamic University, DCS&SE,
Islamabad, Pakistan
[2]University of Tartu, Ulikooli 18, 50090 Tartu. Estonia

Abstract: Cluster Labeling is the process of assigning appropriate and well descriptive titles to text documents. The most suitable label not only explains the central theme of a particular cluster but also provides a means to differentiate it from other clusters in an efficient way. In this paper we proposed a technique for cluster labeling which assigns a generic label to a cluster that may or may not be a part of the text document cluster. It finds the theme of a document and designates it as its label. We used Term Frequency and Inverse Document frequency at baseline for tf-idf, with the Term Frequency calculation refined by using a thesaurus. WordNet was used as an external resource for hypernym generation of the terms having the K-Highest tf-idf. The hypernyms with the highest frequency are then taken as the label of the cluster. The details of the datasets used for experimentation and the comparative results with existing methods are provided in the paper, and clearly reflects the meaningful outcome of our technique.

Keywords: clustering, cluster labeling, WordNet, thesaurus

1. INTRODUCTION

One of the well-known features of text mining is Document Clustering, or the breaking of large text documents into clusters. Users are then able to use these groups of text for analysis and other purposes. The well-defined clusters have high similarity in its inner group items and low similarity with outer group items. Such clustering chunks may be shaped in a more useful way by assigning each of them an appropriate label through a process called Cluster Labeling. Cluster labeling resolves the problem of weak readability [1] and helps users to understand the theme of the cluster. It also assists in checking whether a particular cluster contains the information relating to a particular interest.

This paper presents an automatic approach for cluster labelling. The concept of term frequency dictates that the word which appears the most in a cluster is the one assigned as the label. In cases where this is not possible, our approach allows a generic label that describes the theme of the cluster to be assigned. Objective of assigning label to a cluster is achieved through:

- Finding tf-idf of words appearing in document cluster
- Refinement of words through thesaurus using WORDNET
- Selection of Cluster label using hypernym frequency

The rest of the paper is organized into four

Received, February 2016; Accepted, August 2016
*Corresponding author: E-mail: Syed Muhammad Saqlain; Email: syed.saqlain@iiu.edu.pk

sections. Section 2 includes a review of existing clustering and cluster labelling research. Section 3 is about our proposed methodology while Section 4 discusses the results of our experiment. Section 5 gives our conclusion based on the results and future work directions.

2. LITERATURE REVIEW

Text clustering has been a major area of attention when researching text mining techniques. However, due to the need for and importance of cluster text labeling, researchers have also started to explore techniques related to it. Carmel et al. [2] enhanced the labeling method by using Wikipedia .In this technique documents are initially indexed and then clustered using the well-known clustering techniques. Important terms are extracted from each cluster by using the technique described in Cutting et al. [3].For each important term related Wikipedia pages are extracted. Final labels are selected by the use of pointwise mutual information [4] and statistical co-occurrence [5]. Use of Wikipedia for assigning label to clusters may not be very useful as information provided in Wikipedia pages is not accurate. Any person at their own may provide information without any authenticity whose use for labeling purpose would lead to poor results. Authors didn't mention the cases where Wikipedia cases are not available. Ahmad and Khanum [6] described an algorithm called EROCK (Enhanced Robust Clustering Algorithm for Categorical Attributes)which can make and label clusters. They initially arranged into documents (i.e. clusters) then established the link between each through cosine similarity. Ahmad and Khanum then assigned the word that most frequently appeared as the label. Use of frequent term for the purpose of cluster labeling may not produce the accurate result as a term would not be able to give the central theme of the cluster. Moreover authors have not provided any detail for the cases where more than one terms having same frequency would occur. There is no criteria defined to select amongst them. A document with diverse theme would not be handled accurately using technique under discussion.

Tseng et al. [7] presented a hypernym search algorithm for labeling cluster. The proposed technique creates a generic title based on WordNet. By using correlation coefficients (CC), specific words related to cluster are extracted, while hypernym search algorithm determined the final labels and maps it into WordNet. They used WordNet as an external resource for finding labels of the clusters. They have used hypernyms of all the keywords obtained through CC technique. Refinements of the words are not done. Authors provide no mechanism for selection of label amongst different themes obtained through hypernym search algorithm. Review of some more relevant techniques is provided as well.

Pantel and Ravichandran [8] described a method that automatically assigns labels to clusters based on semantic relationships. They then selected the terms that most describe the clusters as the representative words. Bouras and Tsogkas [9] proposed enhancements in the k-means algorithm which used WordNet before clustering and then labelled the cluster. The authors improved the efficiency as compared to k-means clustering and the quality of labels is improved as well.Carmel et al. [10] presented term extraction in the domain of word cloud generation. They used tag-boost method which boost the terms occasionally used by people to tag the content. They claimed to achieve robustness in compared to other techniques. Mehrotra et al.[11] used unmodified Latent Dirichlet Allocation (LDA) to topic model for short text. They have used Twitter dataset and by using hashtag pooling with LDA achieved improvement compared to unchanged LDA. Morik et al. [12] produced structures for navigating social websites.They considered this as an optimization problem and solved it by using Genetic Algorithm. Jiang [13] provided a survey about information extraction. Majorly survey is about, named relation extraction and entity recognition.

Sun [14] proposed a technique for short text classification by using a non-parametric approach. They selected a small set of words based on their defined criteria and matched it with query words. Authors achieved better classification through this approach. Roitman et al. [15] labelled the clusters using the fusion method. They argued that the label of the cluster should be stable even if there are missing data in the clusters. They tested their technique on different datasets and achieved better performance. Alfred et al. [16] used hierarchical agglomerative clustering for document clustering. The agglomerative clustering was used to counter the fact of other clustering techniques that in most

of the cases in prior number of clusters are unknown. They applied different distance measures to investigate the quality of different clusters. They applied different distance measures to investigate the quality of different clusters. Nayak et al. [17] presented a clustering and cluster labelling method for Wikipedia documents. They took Wikipedia as a subset of the whole web. They tested their technique with 1000, 10000 and 50000 clusters. Xu et al. [18] proposed an approach which achieved Chinese word similarity by using hybrid hierarchical structure through HowNet. They performed their experiment on a SemEval 2012 dataset. Matthias et al. [19] used queries for cluster labeling. They combined internal and differential cluster labeling techniques for acquiring desired results. Daoudet al. [20] presented cluster labeling technique which over the clusters of Arabic tweets. They used key terms as candidate labels and through the web enriched them. They used Bayesian network to find semantic relation between the enriched terms. Hurtado et al. [21] proposed methodology for finding topics from collection of documents. The used association rule based pattern mining for their proposed research. They presented a forecasting technique as well which predicted the recognition of a topic in coming future. Diogo and Jonice [22] used topic labeling technique to detect innovative knowledge from scholar data. They completed their research by performing operations of candidate selected, score ranking and label selection. In [23], Alicante et al. used semantic labeling technique for medical data clusters. They constructed word embedding feature dictionary from Wikipedia pages which was later on used for feature creation and cluster labeling. The clusters were formed using k-means clustering algorithm.

3. PROPOSED METHODOLOGY

The proposed technique selects labels for text clusters in two phases: 1) the pre-labeling phase (may also be called as pre-processing phase) and 2) the labeling phase. In each phase, various steps are performed on text data.

3.1 Pre-Labeling Phase

The pre-labeling phase may also be called as labeling pre-processing phase. Clustering, stemming and term extraction are part of pre-labeling phase. The aim of these steps is to make data clean and eligible for accurate labeling.

3.1.1 Clustering

In this step a given dataset is partitioned into a number of similar homogeneous groups. The clusters are formed using a well-known hierarchical clustering method which combines

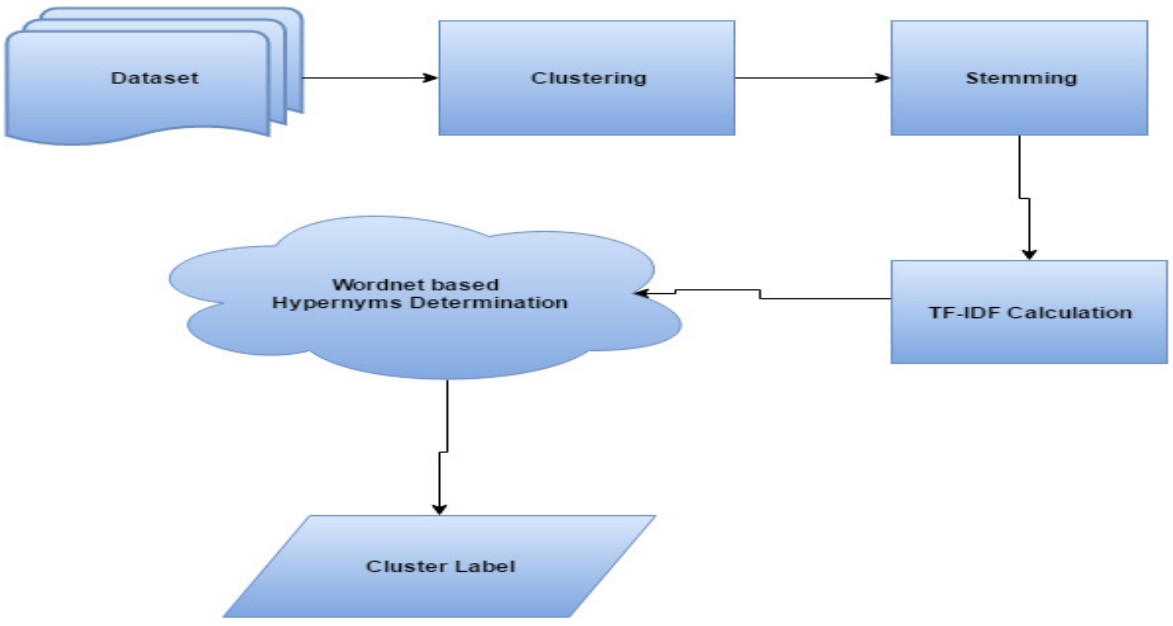

Fig. 1. Flow diagram of proposed technique.

similar observations called agglomerative K-Mean Clustering [10]. Given the document collection D, clustering results in converting D into a set of clusters i.e., C= C_1, C_2,..,C_k. A cluster is represented by its centroid and documents member of each cluster.

3.1.2 Stemming

After the formation of clusters the next phase is stemming. The main aim of this step is to eliminate common words like full stop, commas, articles and other irrelevant words. In order to eliminate common word we make some modification in standard Porter Stemmer algorithm [24] to eliminate not only postfix and prefix but also more and more common words.

Algorithm 1: Stemming

INPUT: Source Blocks SB_X,SB_Y

OUTPUT: Cluster with stemmed words, SC

1: Match $W_i \in C, \forall 1 \leq i \leq n$:n is total no of words in C(Cluster) with $W_p \in CWL, \exists 1 \leq p \leq m$:m is total no of words in CWL(Common Word list)
2: **if** Match is true **then**
3: Remove $W_i$from C
4: **else**
5: Keep $W_i$in C
6: **end if**
7: SC:=C

3.1.3 Term Extraction

In this phase we automatically extract a list of significant terms t(C) = (t_1, t_2, t^1,...,t_n) by calculating the term frequency and inverse document frequency for each word in each given cluster $C_i \in C$ and i= 1,2,3,...k. Term frequency is defined in this situation as the number of times a term appears in a document. Inverse document frequency, on the other hand, is the measure of the general importance of the term based on this formula:

$$idf(t) = \log \frac{\|D\|}{d: t \in d} \quad (1)$$

$$tfidf = tf(t) * idf(t) \quad (2)$$

The selected words are set as candidate words and will be used in the labelling phase (see section 3.2 and Algorithm 3).

Algorithm 2: Term Frequency Calculation using Thesaurus(TFC)

INPUT: SC

OUTPUT: Term Frequency List (TFL)

1: Create a Two dimensional dynamic list TFL initially with size=q×2
2: Initialize p=1
3: Get first word fw from SC and assign
4: TFL[p][1]:=fw
5: TFL[p][2]:=1
6: for each word $W_i \in$ SC
7: **if** W_i matches TFL[j][1] , ∃ j **then**
8: TFL[j][2]:= TFL[j][2]+1
9: **else**
10: p:=p+1
11: TFL[p][1]:=W_i
12: TFL[p][2]:=1
13: **end if**
14: for each list item k in TFL
15: Find thesaurus of TFL[k][1] through WordNet
16: **if** thesaurus(TFL[k][1]) matches some TFL[m][1] **then**
17: TFL[k][2]:= TFL[k][2]+TFL[m][2]
18: Remove TFL[m][ ] from list
19: **end if**

3.2 Labeling Phase

Once the terms are extracted, the Labelling Phase commences. This phase is what we consider as the main step of the proposed technique in which a final label for a particular cluster is assigned or generated. We do this by obtaining the hypernyms of each candidate using WordNet. The hypernyms with the highest frequency is selected as the label of the cluster. The label selected may not

necessarily be a word that can be found in the cluster but a generic label based on the WordNet hypernyms.

Algorithm 3: Cluster Labeling

INPUT: Thresholded Highest TF-IDF Term List (MFTL: Most Frequent Term List)

OUTPUT: Cluster Label (CL)

1: Create a dynamic hypernym list, HL
2: for each potential candidate word,$pcw_j$from MFTL
3: find pcw_j in WordNet
4: **if** match found **then**
5: Add hypernym of pcw_jin HL
6: **else**
7: Add pcw_jin HL
8: **end if**
9: find frequency of each hypernym in HL
10: CL:=hypernym with highest frequency

4. EXPERIMENTAL SETUP, RESULTS AND DISCUSSION

This section describes our experimental setup and the results we obtained after applying our proposed technique. Manual labeling from experts are taken as the perfect labels and their accuracy is bench marked as 100%.Below are the details of each of step.

Data sets: Four different types of text data sets were chosen for the experiment. Table 1 shows details about datasets used for experiments.

Table 1. Dataset details.

Data Set	No of Clusters	Text Chunks
Daily Jang Newspaper	1	2000 Text words/cluster
ODP	6	2000 Text words/cluster
20-News Group	20	2000 Text word/Cluster
Reuter	6	2000 Text words/cluster

In order to evaluate the proposed technique, we have performed experiments over the four datasetsi.e., Daily Jang Newspaper, ODP, Reuter and 20-News Group. Daily Jang Newspaper dataset contains 1 Cluster, ODP has 6, Reuter has 6 and 20-News Group has 20 categories. Each of the dataset has 2000 words per cluster.

Experiment I

In this first data from the Daily Jang newspaper, the document is considered as one cluster containing information about different types of games. It is reduced by applying a modified stemmer algorithm and a term extraction step to pick up top words. Using a thesaurus, the strength of the top words is reduced to greater or equal to threshold level. All of the top words are then mapped using WordNet so that accuracy may be achieved in generating the final cluster label. Results of this data set are shown in Table 2.

In Table 2 details of Experiment I are provided.

Table 2. Detailed Results for Experiment I.

Category/ Cluster	Words having highest tf-idf	Refined Words using thesaurus	Cluster labels through hypernyms
SPORT	1. Football	1. Football	1. A type of sport.
	2. Player	2. Player	2. Sports Man.
	3. Hockey	3. Hockey	3. A type of sport.
	4. Ground	4. Cricket	4. A type of sport.
	5. Cricket	5. Match	
	6. Captain		
	7. Match		

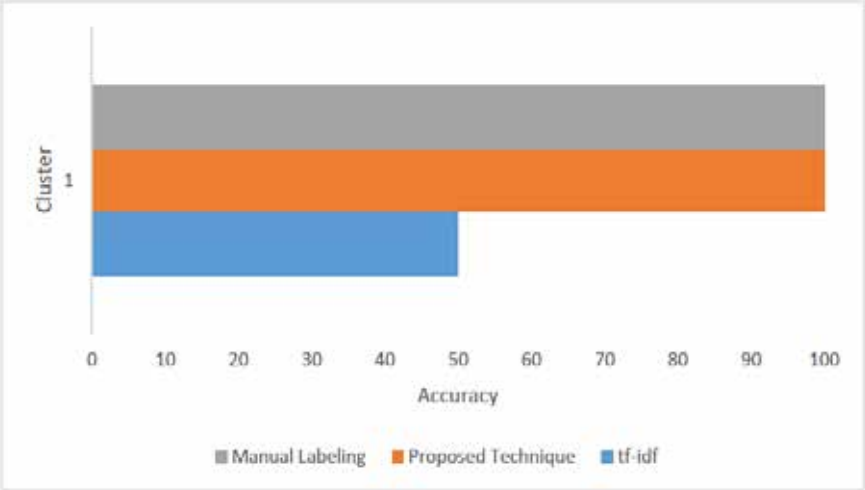

Fig. 2. Comparative results of proposed technique over Jang dataset: Experiment I.

Input dataset contains documents of sports topic. All of these documents are treated as one cluster. Initially Words having highest tf-idf are selected. Important keywords are given as input to algorithm TFC using thesarus where refined words are obtained. Lastly Hypernyms of refined words are obtained through Cluster Labeling algorithm which gave hypernyms,a generic word, of refined words.

Fig. 2 shows comparative results Manual Labeling, tf-idf and the proposed technique in graphical form. Results show that proposed technique achieved the same accuracy as of manual labeling. Performance of proposed technique for Experiment-I is double to tf-idf technique for cluster labeling.

Experiment II

The second collection was gathered by downloading pages from the Open Directory Project (ODP). For this purpose, we randomly

Table 3. Detailed Results for Experiment II.

ID	Category/ Cluster	Words having highest tf-idf	Refined Words using thesaurus	Cluster labels through hypernyms
1	Animals	Rabbit, John, Horse, Cluster, Name, Dog	Rabbit, Horse, Cluster, Dog	1. Herbivorous/Animal 2. Herbivorous/ Animal 3. Group of similar things 4. Carnivores/ Animal
2	Automobile Information		CNG, Fuel, Truck, Car, Automobile, Road	1. A substance 2. A vehicle 3. A vehicle
3	Air Line Information	Column, Scan, New, John, Code	Code, New	1. Unfamiliar, Unknown 2. Rules, Principle, Law
4	Language	John, Claim, Enough, Cluster, Germany	Claim, Enough, Cluster	1. Demand for something 2. Sufficient for something 3. Grouping of similar thing.
5	Male Expectation	Life, Age, Year, Africa, Expectation	Life, Age, Year, Expectation	1. Mode of Living 2. How long something exists. 3. Period of time Expectation
6	Protein Amount	Fat, Protein, Beef, Amount, Calcium	Fat, Protein, Beef, Calcium	1. Bodily Property 2. Substance of Egg 3. Beef Cattle 4. Metallic Item

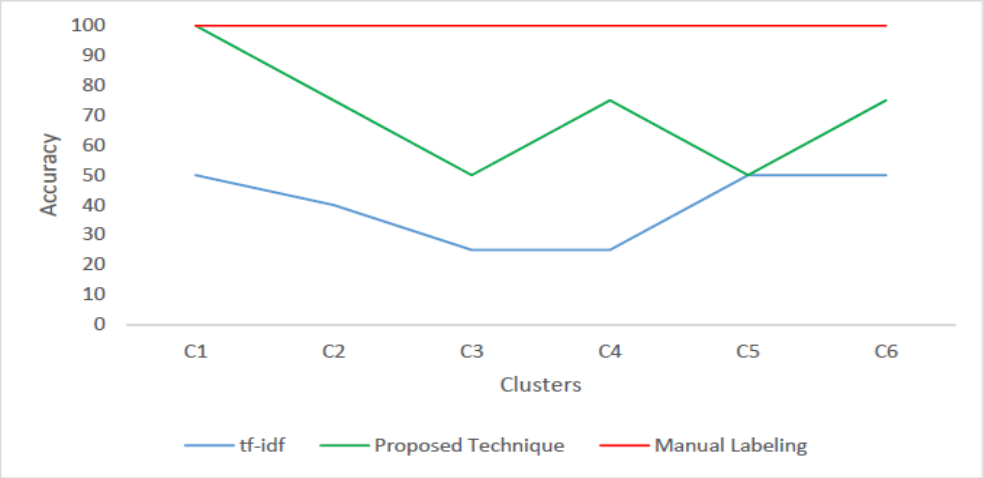

Fig. 3. Comparative results of proposed technique over ODP dataset: Experiment II.

selected six different categories from the ODP hierarchy. For each category we then randomly selected up to 100 documents, resulting in a collection size of about 10,000 documents. We then manually labelled the categories in both collections. These ground truth ”correct” labels were later used to evaluate our labeling system.

Table 3 gives detailed results of Experiment-II performed over proposed technique. 5 highest tf-idf words are selected as label candidate which are subject to TFC using thesaurus algorithm for refinement. Lastly hypernyms are taken as final cluster labels using Cluster Label algorithm.

Fig. 3 depicts comparative results of proposed technique with tf-idf and manual labeling over ODP dataset. For the category of animals our proposed technique produced the same result as by expert human resulting 100% accuracy in comparison to 50% of tf-idf. For the category of automobile information our technique produced A Vehicle as a label whilst tf-idf produced CNG, a less appropriate word. For all the remaining four categories, although proposed technique didn't produce the same label as human expert yet it was more appropriate than produced by tf-idf. Overall proposed technique attained 70% accuracy as compared to 40% through tf-idf over ODP dataset.

Experiment III

Experiment III is performed over dataset having twenty clusters that contain different news stories. We initially stemmed the data set, extracted the candidate terms and then further refined it using a

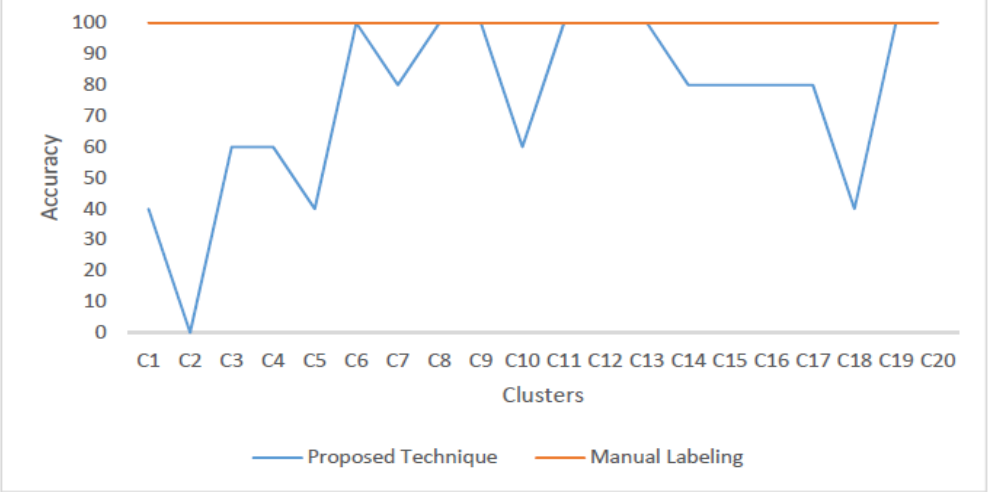

Fig. 4. Comparative results of proposed technique with manual labeling:Experiment III.

thesaurus. In the final step hypernyms of refined words are extracted and mapped to final label. Experiment was performed over 20-News Group dataset which contained approximately 20000 documents from 20 different newspapers. First 5 categories are closely related to each other i.e., computer related documents, for categories 6-9 subject matter is sports related, in categories 11-13 documents contained different topics related to scientific information, whilst category 14 contained documents having forsale topics, documents in categories 15-17 have political talks on three different areas i.e., misc., guns, mildeast. Categories 18-20 have documents containing religious topics in three different areas.

Fig. 4 is the graphical representation of comparative results achieved by proposed technique to expert labeling. Results achieved for first five categories having computer related documents accuracy of the proposed technique is just average. For a particular cluster two having miscellaneous MS windows documents our technique was failed to provide any suitable label. Proposed technique achieved an average 40% accuracy for clusters through 1-5. However for the

Table 4. Detailed results for Experiment IV.

Category/ Cluster	Words having highest tf-idf	Refined Words using thesaurus	Cluster labels through hypernyms
Earn	1. NET 2. QTR 3. Shr 4. Cts 5. Net 6. Revs	1. NET 2. QTR 3. Net 4. Revs	1:goal 2:trap 3:income 4:income
Acquire	1. Acquire 2. Acquisition 3. Stake 4. Company 5. Share	1. Acquire 2. Stake 3. Share	1:device 2:stock certificate, stock 3:wedge
Money	1. Currency 2. Money 3. Market 4. Central banks 5. The Bank 6. Yen	1. Currency 2. Money 3. Market 4. Yen	1:currency 2:currency 3:marketplace, mart 4:China Currency
Grain	1. Wheat 2. Grain 3. Tones 4. Agriculture 5. Corn	1. Wheat 2. Grain 3. Corn	1:seed/ eating food 2:cereal, cereal grass 3:foodstuff, 4:food product
Crude/fuel	1. bpd 2. Crude oil 3. OPEC, 4. mln barrels 5. Petroleum	1. Crude oil 2. OPEC, 3. Petroleum	1:lipid, lipide, lipoid 2:fuel/oil 3:fossil fuel
Trade	1. Trade 2. Tariffs 3. Trading 4. Surplus 5. Deficit 6. Gatt	1. Trade 2. Tariffs 3. Trading 4. Surplus	1:business 2:UN business agency 3:prevailing wind 4:Business rule

clusters 6-9 (i.e., sports related categories)the proposed technique achieved 95% accurate labels. Scientific topics are contained in clusters 10-13 and the proposed technique resulted promisingly by achieving 90% accuracy in labeling clusters. Cluster 14 was about documents having concept of forsale and it is accurately been labeled. Clusters 15-17 are labeled with 80% accuracy whilst labeling accuracy for clusters 18-20 remained 80% as well through proposed technique. Overall accuracy achieved for 20-Newsgroup dataset is 75%.

Experiment IV

In this experiment we have used Reuter-21578 dataset. As originally Reuter dataset has 21578 text documents and multiple categories, we have used largest six categories amongst them for our experimentation i.e., Earn, Acquire, Money, Grain, Fuel and Trade.

Table 4 shows the results obtained through different steps of proposed technique over Reuter dataset. Top five words are selected on the basis of tf-idf and where more than one words have same frequency, all are selected.Results achieved through proposed TFC using thesaurus algorithm and cluster labeling algorithm are presented as well.

Fig. 4 shows comparative results of proposed technique Tseng [7], tf-idf and Manual labeling. For cluster 1 with Earn category, accuracy in cluster labeling for proposed technique matches with accuracy of manual labeling and leading Tseng and tf-idf accuracy. Proposed technique was unable to select appropriate label for cluster 2 and accuracy of proposed technique remained at bottom. For cluster 3 proposed technique achieved same accurate label as tf-idf better than Tseng label and lower than Manual label. As far as cluster 4-6, proposed technique achieved 100% accurate label and leading both tf-idf and Tseng labels. Overall the proposed technique 90% accuracy in labeling clusters of Reuter dataset as compared to 66% accuracy of tf-idf and 72% accurate labels of Tseng.

The results obtained by applying the proposed technique on various text data sets reflect that the performance of the proposed technique is better in terms of its accuracy. After the comparison of the proposed technique with other existing techniques, it is clear that the performance of proposed technique is quite improved. This is also evident in the graphical representation of each experiment. However, the technique proposed do have some constraints. WordNet doesn't cover all the terms extracted from text clusters. Some of the WordNet-generated titles may not also reflect the theme of a particular cluster accurately.

5. CONCLUSION AND FUTURE DIRECTIONS

Cluster labeling is the process of allocating appropriate title to a particular cluster. In our approach, we labelled clusters using an external resource called WordNet. To achieve the task three algorithms are presented Modified Stemming algorithm, Term Frequency Calculation using

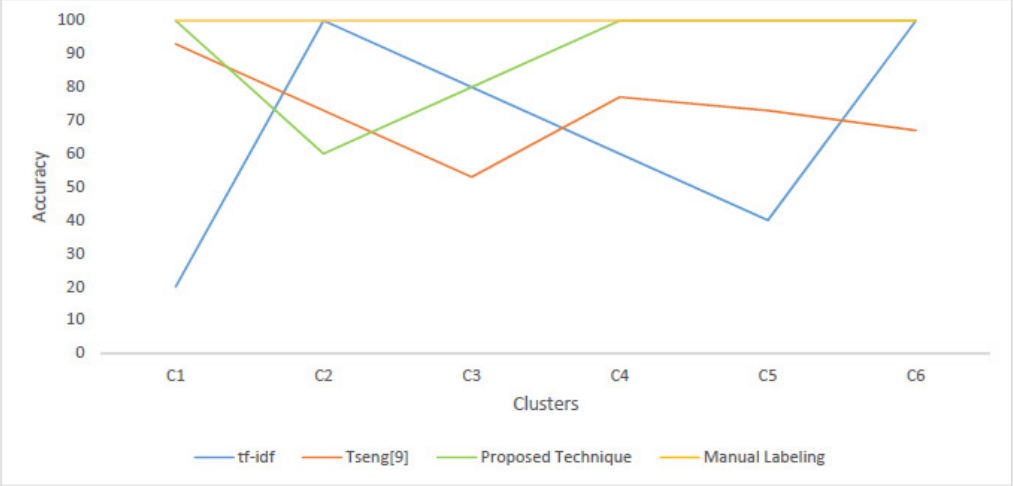

Fig. 5. Comparative results of proposed technique over Reuter dataset.

Thesaurus and Cluster labeling algorithm. TFC and Cluster labeling algorithms use WordNet as an external resource to get Thesaurus and hypernyms. We have performed experiments using four datasets Daily Jang Newspaper, ODP, 20-Newsgroup and Reuter. Experimental results achieved through proposed technique are quite encouraging by attaining 100%, 70%, 75% and 90% accuracy in labeling the clusters for Daily Jang Newspaper, ODP, 20-Newsgroup and Reuter datasets respectively. In different experiments comparative results of proposed technique along with tf-idf, manual labeling and Tseng are presented as well. Comparisons results have clear reflection of achieving better results than Tseng and tf-idf techniques whilst achieved comparable results against manual labeling. Although experimental results reflect that Cluster labeling using WordNet has shown promising results but the performance may be affected by topics whose WordNet hierarchy is not available. We also observed this in those that require multi topic labels. Unfortunately, the proposed system may be unable to generate multi topic label for a particular system. For such collections, there is a need to take an intelligent decision regarding multi topic labeling with the use of WordNet and thesaurus. This could be taken as future direction to improve our proposed technique.

6. REFERENCES

1. Fayyad, U., G. Piatetsky-Shapiro, & P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine* 17: 37-54 (1996).
2. Carmel, D., H. Roitman, & N. Zwerdling. Enhancing cluster labeling using Wikipedia. In: Proceedings of 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA, July 19 – 23, 2009, p. 139–146 (2009).
3. Cutting, D.R., D.R. Karger, J.O. Pedersen, & T.J. Scatter. Gather: A cluster-based approach to browsing large document collections. In: *Proceedings of 15th International ACM SIGIR Conference on Research and Development in Information Retrieval,* Copenhagen, Denmark, June 21–24, 1992, p. 318-329 (1992).
4. Zhang, C. & H. Xu, Clustering Description Extraction Based on Statistical Machine Learning. In: *Proceeding of Intelligent Information Technology Application, IITA'08*, Shanghai, China, December 21–22, 2008, p. 22-26 (2008).
5. Phyu, T.N. Survey of classification techniques in data mining. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, Kowloon, Hong Kong, March 18–20, 2009, p. 18–20 (2009).
6. Ahmad, R. & A. Khanum. Document topic generation in text mining by using cluster analysis with EROCK. *International Journal of Computer Science & Security* 4: 176-182 (2008).
7. Tseng, Y.H., Lin, C.J., Chen, H.H. & Lin, Y.I. Toward generic title generation for clustered documents. In: *Proceedings of Third Asia International Symposium*, Singapore, October16–18, 2006, p. 145-157 (2006).
8. Pantel, P. & D. Ravichandran. Automatically Labeling Semantic Classes. In: Human Language Technology conference / North American chapter of the Association for Computational Linguistics, Boston, USA, May 02–07, 2004, p. 321-328 (2004).
9. Bouras, C. & V. Tsogkas. A clustering technique for news articles using WordNet. *Knowledge-Based Systems* 36: 115-128 (2012).
10. Carmel, D., E. Uziel, I. Guy, Y. Mass, & H. Roitman. Folksonomy-based term extraction for word cloud generation. *ACM Transactions on Intelligent Systems and Technology* 4: 60 (2012).
11. Mehrotra, R., S. Sanner, W. Buntine, & L. Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: *Proceedings of 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 28 – August 01, 2013, p. 889-892 (2013).
12. Morik, K., A. Kaspari, M. Wurst, & M. Skirzynski. Multi-objective frequent termset clustering. *Knowledge and information systems* 3: 715-738 (2012).
13. Jiang, J. Information extraction from text. In: *Mining Text Data*, Charu C. Aggarwal and Cheng Xiang Zhai (Ed). Springer, USA, p. 11–41 (2012).
14. Sun, A. August. Short text classification using very few words. In: *Proceedings 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, USA, August 12–16, 2012, p. 1145-1146 (2012).
15. Roitman, H., S. Hummel, & Shmueli-Scheuer. A fusion approach to cluster labeling. In: *Proceedings 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, Gold Coast, Australia, July 06–11, 2014, p. 883-886 (2014).
16. Alfred, R., T.S. Fun, A. Tahir, C.K. On, & P. Anthony. Concepts labeling of document clusters using a hierarchical agglomerative clustering (hac) technique. In: *Proceedings of the 8th*

International Conference on Knowledge Management in Organizations, Kaohsiung, Taiwan, September 09–13, 2014, p. 263-272 (2014).

17. Nayak, R., R. Mills, C. De-Vries, & S. Geva. Clustering and labeling a web scale document collection using Wikipedia clusters. In: *Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval and Reasoning*, Shanghai, China, November 03, 2014, p. 23-30 (2014).
18. XU, J., LIU, J. and ZHANG, Y. Word Similarity Computing Based on Hybrid Hierarchical Structure by HowNet. Journal of Information Science & Engineering 6: 2089-2101(2015).
19. Hagen, M., M. Michel, & B. Stein. What was the query? Generating queries for document sets with applications in cluster labeling. In: *20th International Conference on Applications of Natural Language to Information Systems*, Passau, Germany, June 17–19, 2015, p. 124-133 (2015).
20. Daoud, D., A. Al-kouz, K. Hasssan, & L. Milliam. Arabic tweets clustering and labeling based on lingual and semantically enriched bayesian network model. *Recent Patents on Computer Science* 8: 112-116 (2015).
21. Hurtado, J.L., A. Agarwal & X. Zhu. Topic discovery and future trend forecasting for texts. *Journal of Big Data* 3: 1-21(2016).
22. Nolasco, D. & J. Oliveira. Detecting knowledge innovation through automatic topic labeling on scholar data. In: *49th Hawaii International Conference on System Sciences (HICSS)*, January 05–08, 2016, Koloa, HI, USA, p. 358-367 (2016).
23. Alicante, A., A. Corazza, F. Isgrò, & S. Silvestri. Semantic cluster labeling for medical relations. In: *4th International KES Conference on Innovation in Medicine and Healthcare*, island of Tenerife, Canary Islands, Spain, June 15–17, 2016, p. 183-193 (2016)
24. Porter, M.F. An algorithm for suffix stripping. *Program* 14: 130-137 (1980).