IADIS International Journal on Computer Science and Information Systems
Vol. 12, No. 1, pp. 14-25
ISSN: 1646-3692

A TEXT SIMILARITY MEASURE FOR DOCUMENT CLASSIFICATION

Gali Suresh Reddy[1] and T. V. Rajinikanth[2]
[1]*Department of Information Technology, VNR VJIET, Hyderabad, India*
[2]*Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, India*

ABSTRACT

Dimensionality reduction is very challenging and important in text mining. We need to know which features be retained what to be and It helps in reducing the processing overhead when performing text classification and text clustering. Another concern in text clustering and text classification is the similarity measure which we choose to find the similarity degree between any two text documents. In this paper, we work towards text clustering and text classification by addressing the use of the proposed similarity measure which is an improved version of our previous measures. This proposed measure is used for supervised and un-supervised learning. The proposed measure overcomes the disadvantages of the existing measures.

KEYWORDS

Feature Selection, Feature Reduction, Clustering, Classification, Dimensionality

1. INTRODUCTION

Text mining may be defined as the field of research which aims at discovering; retrieving the hidden and useful knowledge by carrying out automated analysis of freely available text information and is one of the research fields evolving rapidly from its parent research field information retrieval (Andrew Stranieri, & John Zeleznikow, 2005). Text mining involves various approaches such as extracting text information, identifying and summarizing text, text categorization and clustering. Text Information may be available either in structured form or unstructured form. One of the widely studied data mining algorithms in the text domain is the text clustering. Text clustering may be viewed as an unsupervised learning approach which essentially aims at grouping all the text files which are of similar nature into one category thus separating dissimilar content in to the other groups. In contrast to the text clustering approach, the process of text classification is a supervised learning technique with the class labels known well before. In this paper, we limit our work to text clustering and classification. One common

challenge for clustering is the curse of dimensionality which makes clustering a complex task. The second challenge for text clustering and classification approaches is the sparseness of word distribution. The sparseness of features makes the classification or clustering processes inaccurate, inefficient and thus becomes complex to judge the result. The third challenge is deciding the feature size of the dataset. This is because features which are relevant may be eliminated in the process of noise elimination. Also deciding on the number of clusters possible is also complex and debatable. In this paper, we carry out the dimensionality reduction at two stages. The first stage of dimensionality reduction takes in to the consideration elimination of stop words, stemmed words, followed by computation of tf-idf. The second stage of dimensionality reduction is by the use of singular valued decomposition approach (G. Suresh Reddy, T. V. Rajinikanth, & A. Ananda Rao, 2014; G. Suresh Reddy, A. Ananda Rao, & T.V. Rajinikanth, 2015). This is followed by the use of proposed improved similarity measure w.r.t similarity measure (Suresh Reddy et al., 2014; G. Suresh reddy et al., 2015). The proposed measure is applied to supervised learning process and also for the un-supervised learning process.

2. RELATED WORKS

Text mining spans through various areas and has its applications including recommendation systems, tutoring, web mining, healthcare and medical information systems, marketing, predicting, and telecommunications to specify a few among many applications. The authors (Hussein Hashimi, Alaaeldin Hafez, & Hassan Mathkour, 2015), study and propose various criteria for text mining. These criteria may be used to evaluate the effectiveness of text mining techniques used. This makes the user to choose one among the several available text mining techniques. In (Yannis Haralambous & Philippe Lenca, 2014), the authors use the concept of text item pruning and text enhancing and compare the rank of words with the tf-idf method. Their work also includes studying the importance and extending the use of association rules in the text classification. Association rule mining is playing an important role in text mining and is also widely studied, used and applied by the researchers in text mining community. In (Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, & David Chek Ling Ngo, 2014) authors, discuss the importance of text mining in the predicting and analyzing the market statistics. In short, they perform a systematic survey on the applicability of text mining in market research. In (Sajid Mahmood, Muhammad Shahbaz, & Aziz Guergachi, 2014), the authors work towards finding the negative association rules. Earlier in the past decade, the data mining researchers and market analysts were only interested in finding the dominant positive association rules. In the recent years, significant research is carried out towards finding the set of all possible negative association rules. The major problem with finding negative association rules is the large number of rules which are generated as a result of mining. The negative association rules have important applications in medical data mining, health informatics and predicting the negative behavior of market statistics. In (Wen Zhang, Taketoshi Yoshida, Xijin Tang, & Qing Wang, 2010) the authors use the approach of first finding the frequent items and then using these computed frequent items to perform text clustering. They use the method called “maximum capturing”. With the vast amount of information generating in the recent years, many researchers started coming out with the extensive study and defining various data mining algorithms for finding association rules,

obtaining frequent items or item sets, retrieving closed frequent patterns, finding sequential patterns of user interest (Ning Zhong, Yuefeng Li, &Sheng-Tang Wu, 2010). All these algorithms are not suitable for their use in the field of text mining because of their computational and space complexities. The suitability of these techniques in text mining must be studied in detail and then applied accordingly. One of the important challenges in text mining is handling the problems of misinterpretation and less frequency. An extensive survey on dimensionality reduction techniques is carried in (Fodor, I.K., 2002). The authors discuss the method of principal factor analysis, maximum likelyhood factor analysis and PCA (principal component analysis). A fuzzy approach for clustering features and text classification which involves soft and hard clustering approaches is discussed in (Jung-Yi Jiang, Ren-Jia Liou, & Shie-Jue Lee, 2010). An improved similarity measure overcoming the dis-advantages of conventional similarity measures is discussed in (Yung-Shen Lin, Jung-Yi Jiang, & Shie-Jue Lee, 2014; Shie-Jue Lee, & Jung-Yi Jiang, 2014) their work also involves clustering and classification of text documents. In (Sunghae Jun, Sang-Sung Park, & Dong-Sik Jang, 2014), the concept of support vector machines, SVM is used for document clustering. The other significant findings and research works in the area of text mining include work by the researchers (Sofus A. Macskassy, & Haym Hirsh, 2003; Libiao Zhang, Yuefeng Li, Chao Sun, & Wanvimol Nadee, 2013; Chin Heng Wan, Lam Hong Lee, Rajprasad Rajkumar, & Dino Isa, 2012;Lam Hong Lee, Chin Heng Wan, Rajprasad Rajkumar, & Dino Isa,2012; Dell Zhang & Wee Sun Lee, 2006; Wen Zhang, Taketoshi Yoshida, & Xijin Tang, 2008). In the present work, our idea is to design a similarity measure overcoming dis-advantages in Euclidean, Cosine, Jaccard distance measures (Yung-Shen Lin et al., 2014). The proposed measure considers distribution of features of the global feature set.

Dimensionality is addressed in literature using two methods, Feature reduction and Feature selection (Hawashin, Mansour, & Shadi, 2013). In the feature extraction process also called feature reduction, the high dimensional text documents are projected onto their corresponding low dimensional representation in feature space using algebraic rules and transformations. An application of text mining is discussed in the context of Arabic text in (Nafaa Haffar et.al, 2017; N. Haffar, M. Maraoui & S. Aljawarneh, 2016). Data mining and text mining principles are applied for intrusion detection in (Gunupudi Rajesh Kumar, N. Mangathayaru, & G. Narsimha, 2015; Shadi A. Aljawarneh, Raja A. Moftah, Abdelsalam M. Maatuk, 2016; Gunupudi Rajesh Kumar et.al., 2016; Rajesh Kumar Gunupudi, Mangathayaru Nimmala, Narsimha Gugulothu, & Suresh Reddy, Gali, 2017; Shadi A. Aljawarneh et.al, 2011). Reuse of resources in e-learning systems almost become impossible because of bad indexing. Temporal pattern mining in time stamped temporal databases require similarity functions which can handle supports expressed as vectors. Novel similarity measures for temporal context are proposed in (Vangipuram Radhakrishna, P.V.Kumar, & V.Janaki, 2016; Shadi A. Aljawarneh, Vangipuram Radhakrishna, P.V.Kumar, & V. Janaki, 2017; Vangipuram Radhakrishna, Shadi A. Aljawarneh, Puligadda Veereswara Kumar, & Kim-Kwang Raymond Choo, 2016) and (V. Radhakrishna, P. V. Kumar, V. Janaki & S. Aljawarneh, 2016). (Ons Meddeb, Mohsen Maraoui, & Shadi Aljawarneh, 2016) proposes a new model for AHRS (Arabic hand recognition system). The system is modeled considering preprocessing, segmentation, features extraction, classification and post-processing. Recognizing research trends by mining data obtained using search engines has been studied in (Mohammed R Elkobaisi, Abdelsalam M Maatuk & Shadi Aljawarneh, 2015). A review on gene classification is carried in (Shadi Aljawarneh & Bassam Al-shargabi, 2013).

3. SIMILARITY MEASURE

The problem definition includes defining feature function, feature vector, similarity metric design and validation. Our measure designed is based on the presence-absence of feature considered. So, we have 3 possibilities namely, i^{th} feature is present in both text files, i^{th} feature is absent in both text files and i^{th} feature is present in one of text files.

3.1 Feature Function and Vector

Table 1. Feature Function $f_c< w^{(1m)}, w^{(2m)}>$

w_{1m}	w_{2m}	$f_c< w^{(1m)}, w^{(2m)}>$
absence (0)	absence (0)	-1
absence (0)	presence (1)	1
presence (1)	absence (0)	1
presence (1)	presence (1)	0

Consider Table 1 above. Here, $\mathbf{w^{(1m)}}$ and $\mathbf{w^{(2m)}}$ indicate the presence or absence of m^{th} feature in text files f^i and f^j respectively. The presence of feature is denoted by a value 1 and its absence by 0. The feature function, $f_c< w^{1m},w^{2m}>$ evaluates to 0, 1 or -1. We represent text files F_i and F_j as $F^i = \{w^{i1}, w^{i2}, w^{i3}, w^{i4}, w^{i5}... w^{im}\}$ and $F^j = \{w^{j1}, w^{j2}, w^{j3}, w^{j4}, w^{j5}...w^{jm}\}$. Now, we define generalized feature vector expressed as a function of feature function defined in table.1. Let F^1 and F^2 be any two text files, then the feature vector for these two files is denoted using notation FV^{12} and formally represented as Feature-vector $[F^1, F^2]$, $FV_{12} = [f_c<^{w11} ,^{w21}>, f_c< w^{12} ,w^{22}> ...., f_c< w^{1m},w^{2m}>]$.

We define proposed measure using Equation.1

$$Sim = \frac{(D_f + 1)}{2} \tag{1}$$

with

$$D_f = \frac{\sum_{k=1}^{k=m} D_{nr}(D^{ik}, D^{jk})}{\sum_{k=1}^{k=m} D_{dr}(D^{ik}, D^{jk})} \tag{2}$$

and

$$D_{nr}(D^{ik}, D^{jk}) = \begin{cases} [1 - exp^{-(\frac{1-f_c<w^{ik},w^{jk}>}{\sigma k})^2}] ; f_c < w^{ik}, w^{jk} >= 0 \\ -exp^{-(\frac{1-f_c<w^{ik},w^{jk}>}{\sigma k})^2} \quad ; f_c < w^{ik}, w^{jk} >= 1 \\ 0 \qquad\qquad\qquad\qquad ; f_c < w^{ik}, w^{jk} >= -1 \end{cases} \tag{3}$$

$$D_{dr}(F^{ik}, F^{jk}) = \begin{cases} 1\,;\ f_c < w^{ik}, w^{jk} >\neq -1 \\ 0\,;\ f_c < w^{ik}, w^{jk} >= -1 \end{cases} \quad (4)$$

The ratio of $D_{nr}(F^{ik}, F^{jk})$ and $D_{dr}(F^{ik}, F^{jk})$ denotes average contribution of text features. When D_{nr} and D_{dr} are both evaluated to 0, D_f is fixed to -1. The similarity is bounded between 0 and 1 indicating lower and upper bounds. In our case, threshold value denoted by δ is user defined.

3.2 Properties of Proposed Measure

The Table 2 below lists all properties satisfied by proposed measure.

Table 2. Properties of Proposed Measure

S.No	Property
1	Proposed similarity measure symmetric
2	The distribution of each feature from global feature vector is considered
3	Similarity value varies inversely w.r.t presence-absence features
4	Similarity value varies when both documents have all features and further increases with individual features are widely distributed
5	Similarity is minimum when both documents contain no feature.
6	Similarity is maximum when all features are present.
7	Similarity is average when some features are present and some are absent
8	There is a fixed lower and upper limit for similarity value.
9	The feature presence/absence is important to feature frequency in judging similarity value.

3.3 Remarks

Remark-1: (Property 3, 7, 9)

Let two documents be denoted by F^{pk} and F^{qk} representing k^{th} feature in the text documents F^p and F^q respectively. A value $F^{pk} = 0$ indicates absence of the feature and $F^{pk} = 1$ indicates presence of feature, k. Consider two cases, with$F^{pr} = F^{qr} = 1$ and $F^{pr} = 1$ and $F^{qr} = 0$ (or) $F^{pr} = 0$ and $F^{qr} = 1$. A simple common sense, indicates the similarity value computed for the case-1 must be greater than case-2. Here $D_{dr}(F^{ik}, F^{jk})$remains same for both the situations and this value is denoted as y. However, $D_{nr}(F^{ik}, F^{jk})$ remains different for both situations. Let x denote, $\sum_{k=1,k\neq r}^{k=m} D_{nr}(F^{ik}, F^{jk})$.

Then for the former case, we have expression for $D_f = \frac{x + [1 - e^{-(\frac{1}{\sigma_k})^2}]}{y}$ and $D_f = \frac{x - \lambda e^{-(\frac{1}{\sigma_k})^2}}{y}$ for the latter.Since the value obtained using the expression $(x + e^{-(\frac{1}{\sigma_k})^2}) > (x - e^{-(\frac{1}{\sigma_k})^2})$. Hence, the similarity value for the first case is higher compared to second case.

Remark-2: (Property 5, 8)

Similarity value defined by G_{SIM} has least value when both text documents do not contain non-zero values indicating all features are absent in both documents. Consider, $F_{pr} = 0$ and$F_{qr} = 0$. Here, $\beta(F_{ik}, F_{jk}) = 0$, $\alpha(F_{ik}, F_{jk}) = 0$. In such a case, $G_{SIM} = 0$.

Remark-3: (Property 4, 6, 8)

Similarity value defined by G_{SIM} is maximum, when both documents contain the non-zero values indicating all features are present in both documents. Consider the following situation where $F_{pr} = 1$ and $F_{qr} = 1$. Here, $\beta(F_{ik}, F_{jk}) = 1$, $\alpha(F_{ik}, F_{jk}) = 1$. Then, $D_f = 1$.This gives similarity value as $G_{SIM} = 1$

Remark-4: (Property, 9)

Consider scenario, where$F_{ik} = 10\ and F_{jk} = 0$, we know that documents $F_i and F_j$ are not similar w.r.t k^{th} feature. Similarly, $F_{ik} = 16\ and F_{jk} = 26$ indicates some similarity between these documents. In this case, difference remains same. In such a case, feature frequency value loses its importance. We may hence, conclude k^{th} feature gains more importance in presence-absence situation to k^{th} feature is present in both documents.

Remark-4: (Property 1, 2)

Since, similarity is based on standard deviation of k^{th} feature, but not on any other parameters, the proposed measure is symmetric.

3.4 Analysis of Proposed Measure

3.4.1 Best Case

For best case situation, each feature is present in text files being considered. For sake of analysis, we assume two files as f_1 = {1, 1, 1, 1, 1…..m} and f_2 = {1, 1, 1, 1, 1……..m}.This gives the value of

$$D_f \text{ as } D_f = \frac{\sum_{k=1}^{k=m} D_{nr}(F^{ik}, F^{jk})}{\sum_{k=1}^{k=m} D_{dr}(F^{ik}, F^{jk})} = \frac{m}{m} = 1 .$$

The similarity value for best case is hence evaluated to $Sim = \frac{(D_f+1)}{2} = 1$

3.4.2 Worst Case

In the worst case, f_1 = {0, 0, 0, 0, 0…..… m} and f_2 = {0, 0, 0, 0, 0……..m}. In this case, since both the numerator and denominator of N_{avg} evaluate to 0. So,

$$D_f = \frac{\sum_{k=1}^{k=m} N_{nr}(F^{ik}, F^{jk})}{\sum_{k=1}^{k=m} N_{dr}(F^{ik}, F^{jk})} = -1 \quad \text{and } Sim = \frac{(N_{avg}+1)}{2} = 0.$$

3.4.3 Average Case

We divide this situation in to two situations. The first is the worst situation in average case and second includes average case in general. For average case situation, we have presence-absence

combination of features. For sake of analysis; we assume f_1 = {0, 1, 0, 1, 1…..m} and f_2 = {1, 0, 1, 0, 0……..m}. In this case, value of D_f is $D_f = \frac{\sum_{k=1}^{k=m} N_{nr}(F^{ik},F^{jk})}{\sum_{k=1}^{k=m} N_{dr}(F^{ik},F^{jk})} = -1$. The similarity is $Sim = \frac{(D_f+1)}{2} = 0$.

4. CASE STUDY

Due to space constraints we only outline the important stages in computation process through sample values. Consider the document-feature matrix in Table 3. For making the discussion simple, we choose 9 text documents and 10 features obtained after preprocessing phase.

Table 3. Matrix in Frequency Form

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
File -1	0	1	0	0	1	1	0	0	0	1
File-2	0	0	0	0	0	2	1	1	0	0
File-3	0	0	0	0	0	0	1	0	0	0
File-4	0	0	1	0	2	1	2	1	0	1
File-5	0	0	0	1	0	1	0	0	1	0
File-6	2	1	1	0	0	1	0	0	1	0
File-7	3	2	1	3	0	1	0	1	1	0
File-8	1	0	1	1	0	1	0	0	0	0
File-9	1	1	1	1	0	0	0	0	0	0

All these features together represent the global feature set over which the document-feature matrix is formed. This global word set is obtained after initial preprocessing phase. We maintain the matrix in both the frequency form and binary form. After applying SVD decomposition, we get the reduced matrix as shown in Table 4 and Table 5

We compute the standard deviation for the most significant words of the original global feature set instead of all the words in global feature set. This computed standard deviation for each word as shown in Table 6 is later used when computing the similarity degree between any two text files by using the proposed measure. The standard deviation of each word represents the statistical distribution of the corresponding word. Table 7 shows computations of α while computations of beta are straight forward.

Table 4. Reduced Matrix in Frequency Form

	W_6	**W_3**	**W_1**	**W_2**	**W_3**	**W_9**
File1	1	0	0	1	0	0
File2	2	0	0	0	0	0
File3	0	0	0	0	0	0
File4	1	1	0	0	0	0
File5	1	0	0	0	1	1
File6	1	1	2	1	0	1
File7	1	1	3	2	3	1
File8	1	1	1	0	1	0
File9	0	1	1	1	1	0

Table 5. Reduced Matrix in Binary Form

	W_6	W_3	W_1	W_2	W_3	W_9
File1	1	0	0	1	0	0
File2	1	0	0	0	0	0
File3	0	0	0	0	0	0
File4	1	1	0	0	0	0
File5	1	0	0	0	1	1
File6	1	1	1	1	0	1
File7	1	1	1	1	1	1
File8	1	1	1	0	1	0
File9	0	1	1	1	1	0

Table 6. Standard Deviation Matrix

S.Dev	W_6	W_3	W_1	W_2	W_4	W_9	W_8
	0.35	0.53	0.52	0.52	0.52	0.52	0.52

Table 7. Sample Computations Distributions

α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}	α_{18}	α_{19}
0.95	0.97	0.92	0.93	1.92	1.87	0.89	0.92

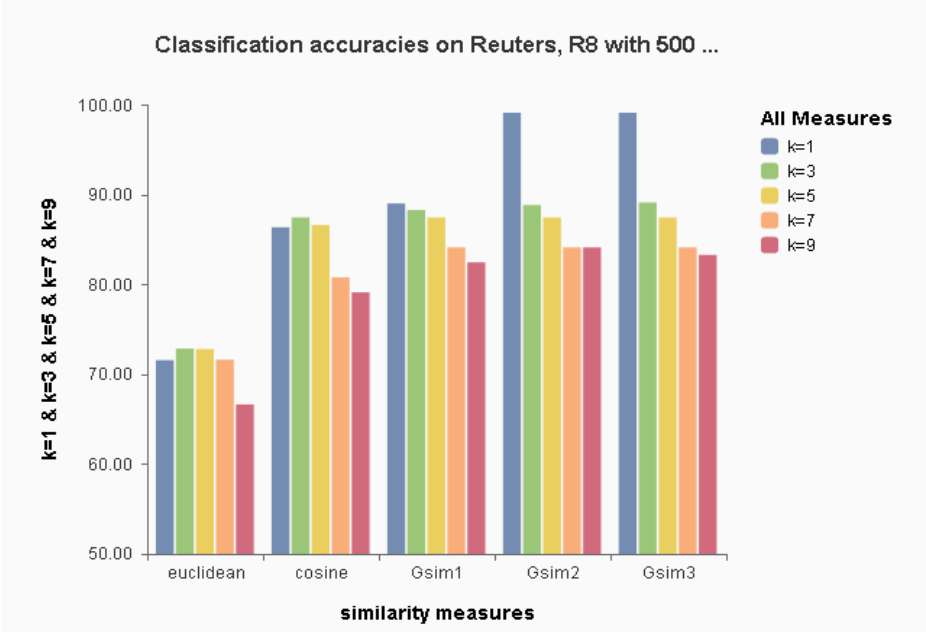

Figure 1. Column Chart Representing Classification Accuracies for k=1, 3, 5, 7, 9

5. RESULTS AND DISCUSSIONS

The figure 1 and figure 2 below shows the results of proposed similarity measure, Gsim-3w.r.t other similarity measures on Reuters, R52 and R8 dataset. The Gsim-1 and Gsim-2 are similarity measures of our previous works (Suresh Reddy et al., 2014)(Suresh Reddy et al., 2015).

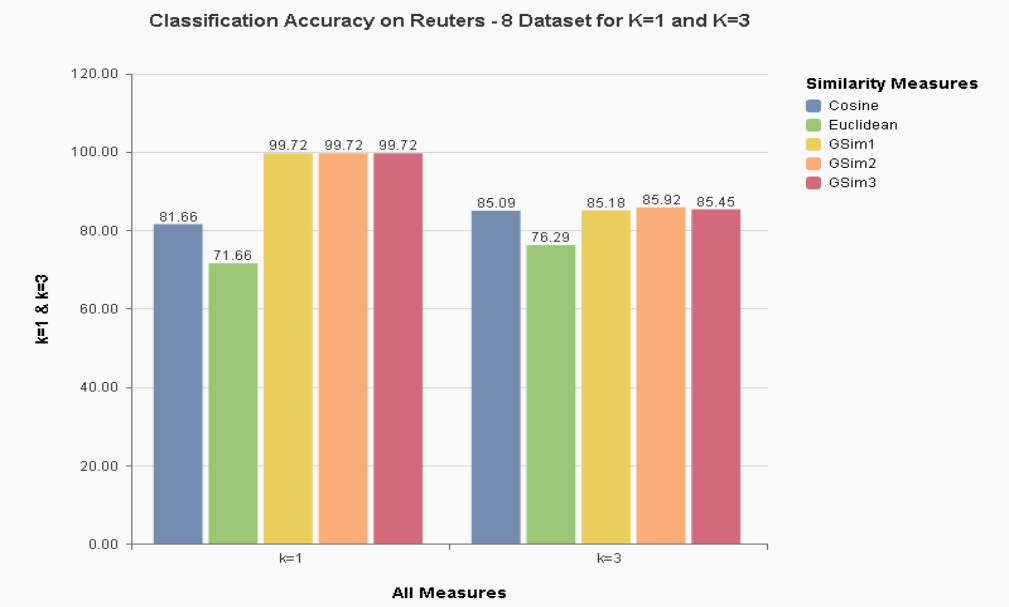

Figure 2. Classification Accuracy on Reuters-8 for k=1 and k=3

6. CONCLUSIONS

In this work, we apply singular value decomposition to perform dimensionality reduction and use reduced dimensionality documents to perform text classification. To perform text classification, we use proposed distance measure which is the improved version of our previous measures which also considers the distribution of features of the document. For text classification, we use the proposed similarity measure and classify the new text document to the corresponding class label of training dataset. The proposed measure is designed by considering, worst case, average and best case situations and validated formally. The results show proposed measure outperforms existing measures.

ACKNOWLEDGEMENTS

This work has been funded under UGC minor research project with File no: 4-4/2014-15(MRP-SEM/UGC-SERO). The principal investigator T.V.Rajinikanth and Co-Principal Investigator, G.Suresh Reddy would like to thank UGC for providing the research project grant.

REFERENCES

Andrew Stranieri, &John Zeleznikow.(2005). Information retrieval and text mining. *Knowledge Discovery from Legal Databases Law and Philosophy Library*, *69*,147-169.

Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, & David Chek Ling Ngo.(2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications,41*, 7653–7670.

Chin Heng Wan, Lam Hong Lee, Rajprasad Rajkumar, & Dino Isa. (2012). A hybrid text classification approach with low dependency on parameter by integrating k-nearest neighbor and support vector machine. *Expert Systems with Applications*, *39*(15), 11880-11888.

Dell Zhang & Wee Sun Lee.(2006). Extracting key-substring-group features for text classification. *In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '06).* ACM, New York, NY, USA, 474-483, doi:10.1145/1150402.1150455

Fodor, I.K.(2002). A survey of dimension reduction techniques. *Technical Report, UCRL-ID-148494, Lawrence Livermore National Laboratory*, Livermore. doi:10.2172/15002155

G. Suresh Reddy, A. Ananda Rao, &T.V.Rajinikanth.(2015). An improved similarity measure for text clustering and classification. *Advanced Science Letters*, *21*(11), 3583-90.

G. Suresh Reddy, T. V. Rajinikanth, & A. Ananda Rao.(2014, July). Design and analysis of novel similarity measure for clustering and classification of high dimensional text documents. *In Proceedings of the 15th International Conference on Computer Systems and Technologies (CompSysTech '14)*, 194-201, doi:10.1145/2659532.2659615

Gunupudi, R. K., Mangathayaru, N., & Narsimha, G. (2015). A novel similarity measure for intrusion detection using gaussian function. *Revista Tecnica De La Facultad De Ingenieria Universidad Del Zulia*, *39*(2), 173-183.

Gunupudi, R. K., Mangathayaru, N., & Narsimha, G. (2016). An approach for intrusion detection using novel gaussian based kernel function. *Journal of Universal Computer Science, 22*(4), 589-604

Gunupudi, R. K., Mangathayaru, N., & Narsimha, G. (2016). Intrusion detection a text mining based approach. *Special issue on Computing Applications and Data Mining International Journal of Computer Science and Information Security (IJCSIS)*, 14, 76-88.

Gunupudi, R. K., Mangathayaru, N., & Narsimha, G. (2015). Intrusion detection using text processing techniques: a recent survey. *Proceedings of the International Conference on Engineering & MIS*. doi:10.1145/2832987.2833067

Hawashin, Bilal; Mansour, Ayman; Aljawarneh, Shadi. (2013). An efficient feature selection method for arabic text classification. *International Journal of Computer Applications*, *83*(17), 1-6.

Hussein Hashimi, Alaaeldin Hafez, & Hassan Mathkour. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior*, *51*, 729-733.

Jung-Yi Jiang, Ren-Jia Liou, & Shie-Jue Lee.(2010). A fuzzy self-constructing feature clustering algorithm for text classification. *IEEE Transactions on Know-ledge and Data Engineering,23*(3), 335 – 349.

Lam Hong Lee, Chin Heng Wan, Rajprasad Rajkumar, & Dino Isa.(2012). An enhanced support vector machine classification framework by using Euclidean distance function for text document categorization. *Applied Intelligence, 37*(1), 80-99.

Libiao Zhang, Yuefeng Li, Chao Sun, & Wanvimol Nadee. (2013). Rough set based approach to text classification. *In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, IEEE Computer Society, Washington, DC, USA, *03*, 245-252. doi:10.1109/WI-IAT.2013.190

Mohammed R Elkobaisi, Abdelsalam M Maatuk, Shadi Aljawarneh. (2015, Aug). A Proposed Method to Recognize the Research Trends using Web-based Search Engines. Paper presented at the International Conference on Engineering & MIS, Istanbul, Turkey. Retrieved from http://dl.acm.org/citation.cfm?id=2833012

N. Haffar, M. Maraoui & S. Aljawarneh. (2016). Use of indexed Arabic text in e-learning system. Paper presented at the International Conference on Engineering & MIS (ICEMIS), Agadir. Abstract retrieved from http://ieeexplore.ieee.org/abstract/document/7745321/

Nafaa Haffar, Mohsen Maraoui, Shadi Aljawarneh, Mohammed Bouhorma, Abdallah Altahan Alnuaimi, Bilal Hawashin. (2017). Pedagogical indexed Arabic text in cloud e-learning system. International Journal of Cloud Applications and Computing, 7(1), 32-46.

Ning Zhong, Yuefeng Li, & Sheng-Tang Wu. (2010).Effective pattern discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering. 24*(1), 30 – 44.

Ons Meddeb, Mohsen Maraoui, & Shadi Aljawarneh. (2016, Sep). Hybrid modeling of an Offline Arabic Handwriting Recognition System AHRS. *Paper presented at the International Conference on Engineering & MIS*, Agadir. Retrieved from http://ieeexplore.ieee.org/abstract/document/7745319/

Reddy, G.S, &Rajinikanth, T.V., & Anandarao, Rao.(2014). A frequent term based text clustering approach using novel similarity measure. *Proceedings of 2014 IEEE International Advance Computing Conference*, 495-499, doi:10.1109/IAdCC.2014.6779374

Rajesh Kumar Gunupudi, Mangathayaru Nimmala, Narsimha Gugulothu, & Suresh Reddy, Gali. (2017). CLAPP: A self constructing feature clustering approach for anomaly detection. *Future Generation Computer Systems*. doi:10.1016/j.future.2016.12.040

Sajid Mahmood, Muhammad Shahbaz, & Aziz Guergachi.(2014). Negative and positive association rules mining from text using frequent and infrequent itemsets. *The Scientific World Journal*, *2014*, Article ID 973750, 11 pages.

Shadi A Aljawarneh, Mohammed R Elkobaisi, Abdelsalam M Maatuk. (2016). A new agent approach for recognizing research trends in wearable systems. *Computers & Electrical Engineering*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0045790616309995.

Shadi A. Aljawarneh, Vangipuram Radhakrishna, P.V.Kumar, V. Janaki. (2017). G-SPAMINE: An approach to discover temporal association patterns and trends in internet of things. Future Generation Computer Systems, http://dx.doi.org/10.1016/j.future.2017.01.013.

Shadi A. Aljawarneh, Raja A. Moftah, Abdelsalam M. Maatuk. (2016). Investigations of automatic methods for detecting the polymorphic worms signatures. Future Generation Computer Systems, 60, 67-77.

Shadi Aljawarneh. (2011). Cloud Security Engineering: Avoiding security threats the right way. International Journal of Cloud Applications and Computing (IJCAC), 1(2), 64-70.

Shadi Aljawarneh & Bassam Al-shargabi. (2013). Gene Classification: A Review. Paper presented at the 6th International Conference on Information Technology. Retrieved from http://sce.zuj.edu.jo/ICIT13/images/Camera%20Ready/Sorftware%20Engineering/650_shadi.pdf

Shie-Jue Lee, & Jung-Yi Jiang. (2014). Multilabel text categorization based on fuzzy relevance clustering. *IEEE Transactions on Fuzzy Systems, 22*(6), 1457-1471.

Sofus A. Macskassy & Haym Hirsh.(2003). Adding numbers to text classification. *In Proceedings of the twelfth international conference on Information and knowledge management (CIKM '03).* ACM, New York, NY, USA, 240-246, doi:10.1145/956863.956910

Sunghae Jun, Sang-Sung Park, & Dong-Sik Jang.(2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness. *Expert Systems and Applications, 41*(7),3204-3212.

V. Radhakrishna, P. V. Kumar, V. Janaki & S. Aljawarneh. (2016). A computationally efficient approach for temporal pattern mining in IoT. *Proceedings of the International Conference on Engineering & MIS,* 1-4. doi: 10.1109/ICEMIS.2016.7745354.

V. Radhakrishna, P. V. Kumar, V. Janaki & S. Aljawarneh. (2016). A similarity measure for outlier detection in time stamped temporal databases. Proceedings of the International Conference on Engineering & MIS, 1-5.doi: 10.1109/ICEMIS.2016.7745347

Vangipuram Radhakrishna, Shadi A. Aljawarneh, Puligadda Veereswara Kumar, Kim-Kwang Raymond Choo. (2016). A novel fuzzy gaussian-based dissimilarity measure for discovering similarity temporal association patterns. *Soft Computing.* First Online: 18 November 2016. doi:10.1007/s00500-016-2445-y

Wen Zhang, Taketoshi Yoshida, & Xijin Tang.(2008). Text classification based on multi-word with support vector machine. *Know. Based Syst, 21*(8), 879-886.

Wen Zhang, Taketoshi Yoshida, Xijin Tang, &Qing Wang.(2010). Text clustering using frequent item sets. *Knowledge-Based Systems*, *23*(5), 379–388.

Yannis Haralambous & Philippe Lenca. (2014).Text classification using association rules, dependency pruning and hyperonymization. *Proceedings of DMNLP, Workshop at ECML/PKDD*,Aachen, Germany, 1202,65-80

Yung-Shen Lin, Jung-Yi Jiang, & Shie-Jue Lee.(2014). A similarity measure for text classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*,*26*(7), 575-1590.