ISSN:0975-9646

Effectiveness of Different Similarity Measures for Text Classification and Clustering

Komal Maher[1], Madhuri S. Joshi[2]

Computer Science and Engineering
*Jawaharlal Nehru Engineering College,*Aurangabad-431003, M.S., India

***Abstract* – Present days humans are associated with large amount of data on regular basis. The sole purpose of generated data is to meet the immediate needs and no attempt in organizing the data for later efficient retrieval. Data mining is a concept of extracting knowledge from such an enormous amount of data.There are many techniques to classify and cluster the data which exists in the structured format, based on similarity between documents in the text processing field.**

Clustering algorithms require a metric to quantify how different two given documents are.This difference is often measured by some distance measure such as Euclidean distance, Cosine similarity, Jaccard correlation, Similarity measure for text processing to name a few. In this research work, we experiment with Euclidean distance, Cosine similarity and Similarity measure for text processing distance measures. The effectiveness of these three measures is evaluated on a real-world data set for text classification and clustering problems. The results show that the performance obtained by the Similarity measure for text processing measure is better than that achieved by other measures.

Keywords* – *Document classification; document clustering; entropy; accuracy; classifiers; clustering algorithms

I. INTRODUCTION

Text processing is a burgeoning new technology for discovery of knowledge. Text processing plays an important role in data mining, and information retrieval [11], [12], [13].Text processing and text mining extended the data mining approach to textual data and is responsible for finding useful and interesting patterns, models, and rules from unstructured texts. In text processing, the bag-of-words model is commonly used [14], [15], [16].

In text mining, a document is represented as a vector in which each component indicates the value of its corresponding feature in the document. The feature value can be the number of occurrences of a term appearing in the document (term frequency), the ratio between the term frequency and the total number of occurrences of all the terms in the document set (relative term frequency), or a combination of term frequency and inverse document frequency (TFIDF) [28]. Usually, most of the feature values in the vector are zero, such high dimensionality and sparsity can be a major challenge for similarity measure which is an important operation in text processing algorithms [17], [18], [19], [20], [22].

A variety of similarity measures have been proposed and widely applied in literature, such as cosine similarity and the Jaccard correlation coefficient. Meanwhile, similarity is often apprehended in terms of distance or dissimilarity as well [23]. Measures such as Euclidean distance and relative entropy have been applied in clustering to calculate the pair-wise distances.

Given the diversity of similarity and distance measures available, their effectiveness in text document clustering is still not clear. Although Strehl et al. compared the effectiveness of a number of measures [24], our experiments extended their work by including more measures and experimental datasets, such as the averaged Kullback-Leibler divergence, which has shown its effectiveness in clustering text and attracted considerable research interest recently. More specifically, we evaluated three measures with empirical experiments: Euclidean distance, cosine similarity, and SMTP (Similarity Measure for Text Processing) distance measure. In order to come up at a sound conclusion we have performed an empirical evaluation with real world data sets that each has different characteristics.

The rest of this paper is organized as follows. The next section briefly describes the related work. Section 3 discusses the Euclidean distance, Cosine similarity and SMTP distance measures and their semantics. Experimental results are presented in Section 4. Finally, concluding remarks are given Section 5.

II. RELATED WORK

A lot of measures have been proposed for computing the similarity between two documents. The survey of existing approaches for finding similarity between two documents has been done after a systematic review with principled approach in which major digital libraries for computer science have been searched. We focused on papers since last 10 years.

Yung Shen Lin et. al. [1]have presented a novel SMTP similarity measure between two documents by embedding several properties in this measure. The proposed scheme has been extended to measure the similarity between two sets of documents. To improve the efficiency, an approximation has been used and the complexity involved in the computation has reduced. The effectiveness of proposed measure has investigated by applying it in k-NNbased single-label classification; k-NN based multi-label classification, k-means clustering, and Hierarchical Agglomerative Clustering (HAC) on several real-world data sets.The results have shown that the performance obtained by the similarity measure for text processing (SMTP) is better than that achieved by other measures.

TABLE I. SUMMARY OF LITERATURE SURVEY

Author Name	Dataset Used	Clustering Technique Used	Similarity Measure Used
Yung Shen Lin, Jung yi Jiang, Shin Jue Lee, 2014	WebKb, Reuters-8, RCV1	HAC, K-means	SMTP
Kalaivendhan.K, Sumathi. P, 2014	-	HAC	Cosine
B Sindhiya and N Tajunisha, 2014	WebKb	HAC, K-means	Concept based SMTP
Pranjal Singh, Mohit Sharma, 2013	WebKb, Wap,Classic	K-mean	Euclidean, Cosine, Jaccard
K Shruti, B Reddy, 2013	-	HAC	Multi- viewpoint based similarity measure
VenkataGopalaRao, S. Bhanu Prasad A, 2013	7 different datasets	K-means	Cosine
P. Sowmya Lakshmi, V. Sushma, T. Manasa, 2012	Reuters-21578	KNN	Eucliden, Manhattan
Anil Kumar Patidar, 2012	KDD cup'99, Mashroom	Shared Nearest Neighbour	Euclidean, Cosine, Jaccard, Person correlation distance
MannanGoyal, NehaAgewal,Manoj Sharma, NayanKalita, 2012	Unstructured dataset	K-means	Cosine, fuzzy
Muhammad Rafi, 2011	News20, Webkb,Classic	K-means	Topic map based similarity measure
Hung Chim, Xiaotie Deng, 2008	News20	HAC	Phrase based similarity measure
Anna Huang, 2008	20news, Webkb,Classic, Hitech, Re0, Tr41, Wap	K-means	Euclidean, Cosine, Jaccard, Person correlation distance
Hung Chim, Xiaotie Deng, 2007	OHSUMED,RCV1	HAC	Suffix tree similarity measure

The concept and term based model represents document as a two-way model with the aid of WordNet. In the two-way representation model, the term information is represented first, and theconcept information is represented second and these levels are connected by the semantic relatedness between terms and concepts.Experimental results presented by B. Sindhiyaet. al.[4] have shown that the proposed model and classification framework significantly improved the performance of classification and clustering by comparing with the existing SMTP model. The experiments also shows that CSMTP(concept and term based similarity measure for text processing)takes less time when running in parallel, less space when running in series and high categorization accuracy.

Kalaivendhan K. et. al. [3] presented HAC and Correlation similarity techniques which are used for any type of text document to display the most relevant document of the clusters. The results opt-out with a conclusion that correlation similarity and HAC algorithm makes similarity and document retrieval more accurate than the cosine similarity and MVS algorithm.The methodology of cluster analysis involved in the study by Pranjal Singh et. al. is evidently partitional and require a similarity measure. The three components that affect the final results are representation of the objects, distance or similarity measures, and the clustering algorithm itself.

Ms.K.Sruthiet. al.[5] introduced multi-viewpoint based similarity measure and related clustering methods for text data. Using multiple viewpoints, more informative assessment of similarity could be achieved and performance is much better than Euclidean, Jaccard or Pearson coefficient similarity measures.Future work tends to explore how they work on other types of sparse and high dimensional data.

VenkataGopalaRao S.et. al. [6]found that except for the Euclidean distance measure, the other measures have comparable effectiveness for the partitioned text document clustering task. Pearson correlation coefficient and the averaged measures are slightly better in that their resulting clustering solutions are more balanced and have a closer match with the manually created category structure.

Neepa Shah has explained the document clustering procedure with feature selection, TFIDF process, dimension reduction mechanisms etc and various improvements in it. In this survey paper, applications, challenges, similarity measures and evaluation of document clustering algorithms is summarized.The paper by

P.Sowmya Lakshmi et. al. [7] attempts to classify the data by employing different similarity measures, with different vector generation technique.

Anil Kumar Patidaret. al. [8]have analyzed the impact upon SNN clustering approach (SNN) of different similarity computation functions and compared the resultant similarity graphs and clusters, which inferred that the SNN clustering approach with Euclidean similarity measure vides better and faster results as compared to the other distance functions.

Manan Mohan Goyalet. al. tried to compare the cosine and fuzzy similarity measure using the k-means algorithm. Muhammad Rafiet. al. proposed the topic map based similarity measure, which is quite effective in clustering documents collection, as it produced more coherent clustering as compared with human categorized structures.

The work by Hung Chimet. al. [10] has presented a successful approach to extend the usage of TFIDF weighting scheme: the term TFIDF weighting scheme is suitable for evaluating the importance of not only the keywords but also the phrases in document clustering.

III. SIMILARITY MEASURES

Before clustering, a similarity or distance measure must be determined. This measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. In many cases, these characteristics are dependent on the data or the problem context at hand, and there is no measure that is universally best for all kinds of clustering problems. Moreover, choosing an appropriate similarity measure is also crucial for cluster analysis, especially for a particular type of clustering algorithms. For example, the density-based clustering algorithms, such as DBScan [33], rely heavily on the similarity computation. Density-based clustering finds clusters as dense areas in the data set, and the density of a given point is in turn estimated as the closeness of the corresponding data object to its neighboring objects. Recalling that closeness is quantified as the distance or similarity value, we can see that large number of distance or similarity computations are required for finding dense areas and estimate cluster assignment of new data objects. Therefore, understanding the effectiveness of different measures is of great importance in helping to choose the best one.

In general, similarity or distance measures map the distance or similarity between the symbolic descriptions of two objects into a single numeric value, which depends on two factors— the properties of the two objects and the measure itself.

A. Metric

Not every distance measure is a metric. To qualify as a metric, a measure d must satisfy the following four conditions.Let x and y be any two objects in a set and d(x, y) be the distance between x and y.

1. The distance between any two points must be nonnegative, that is, $d(x, y) \geq 0$.

2. The distance between two objects must be zero if and only if the two objects are identical, that is, $d(x, y) = 0$ if and only if $x = y$.

3. Distance must be symmetric, that is, distance from x to y is the same as the distance from y to x, i.e. $d(x, y) = d(y, x)$.

4. The measure must satisfy the triangle inequality, which is $d(x, z) \leq d(x, y) + d(y, z)$.

B. Euclidean Distance

Euclidean distance is a standard metric for geometrical problems. It is the ordinary distance between two points and can be easily measured with a ruler in two- or three-dimensional space. Euclidean distance is widely used in clustering problems, including clustering text. It satisfies all the above four conditions and therefore is a true metric. It is also the default distance measure used with the K-means algorithm.

Measuring distance between text documents, given two documents d_a and d_b represented by their term vectors $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$ respectively, the Euclidean distance of the two documents is defined as

$$D_E(\overrightarrow{t_a}, \overrightarrow{t_b}) = \left(\sum_{t=1}^{m} |W_{t,a} - W_{t,b}|^2\right)^{1/2} \quad (1)$$

Where the term set is $T = \{t_1, \ldots, t_m\}$. As mentioned previously, we use the *tfidf* value as term weights, that is $w_{t,a} = tfidf(d_a, t)$.

C. Cosine Similarity

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications [35] and clustering too [34].

Given two documents $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$ respectively, their cosine similarity is,

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|} \quad (2)$$

Where $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$ are m-dimensional vectors over the term set $T = \{t_1, \ldots, t_m\}$. Each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between [0, 1].

An important property of the cosine similarity is its independence of document length. For example, combining two identical copies of a document *d* to get a new pseudo document *d'*, the cosine similarity between *d* and *d'* is 1, which means that these two documents are regarded to be identical.

Meanwhile, given another document *l*, *d* and *d'*will have the same similarity value to *l*, that is, $\text{sim}(\overrightarrow{t_d}, \overrightarrow{t_l})$= $\text{sim}(\overrightarrow{t_d}, \overrightarrow{t_l})$. In other words, documents with the same composition but different totals will be treated identically. Strictly speaking, this does not satisfy the second condition of a metric, because after all the combination of two copies is a different object from the original document. However, in practice, when the term vectors are normalized to a unit length such as 1, and in this case the representation of *d* and *d'* is the same.

D. Similarity Measure for Text Processing

Based on the preferable properties mentioned above, a similarity measure, called SMTP (Similarity Measure for Text Processing), for two documents d1 = <d11, d12, ... , d1m> and d2 = <d21, d22, ... , d2m> defines a function F as follows:

$$F(d_1, d_2) = \frac{\sum_{j=1}^{m} N_*(d_{1j}, d_{2j})}{\sum_{j=1}^{m} N_{\cup}(d_{1j}, d_{2j})} \quad (3)$$

Where,

$$N_*(d_{1j}, d_{2j}) = \begin{cases} 0.5\left(1 + exp\left\{-\left(\frac{d_{1j}-d_{2j}}{\sigma_j}\right)^2\right\}\right), if\, d_{1j}, d_{2j} > 0 \\ 0, if\, d_{1j} = 0\ and\ d_{2j} = 0 \\ -\lambda, otherwise, \end{cases} \quad (4)$$

$$N_{\cup}(d_{1j}, d_{2j}) = \begin{cases} 0, if\, d_{1j} = 0\ and\ d_{2j} = 0 \\ 1, otherwise. \end{cases} \quad (5)$$

Then the similarity measure, S_{SMTP}, for $d_1$and$d_2$is

$$S_{SMTP}(d_1, d_2) = \frac{F(d_1,d_2) + \lambda}{1 + \lambda} \quad (6)$$

This measure takes into account the following three cases: a) The feature considered appears in both documents, b) the feature considered appears in only one document, and c) the feature considered appears in none of the documents.

IV. EXPERIMENTAL RESULTS

In this section, the effectiveness of Euclidean distance; Cosine similarity and SMTP distance measures have been investigated. This investigation is done by applying our measures in couple of text applications, namely, k-NN based classification [26], Naïve Bayes classification and k-means clustering [25]. We compare the performance of SMTP with that of other two measures, Euclidean [27], Cosine [28] in this Section.

Three data sets, named WebKB [30], Reuters-8 [29], and RCV1 [31], respectively, are used in the experiments we followed.

A. Classification Dataset

The randomly selected training documents are used for training or validation and the testing documents are used for testing. Whereas, the data for training or validation are separate from the data for testing in each case.

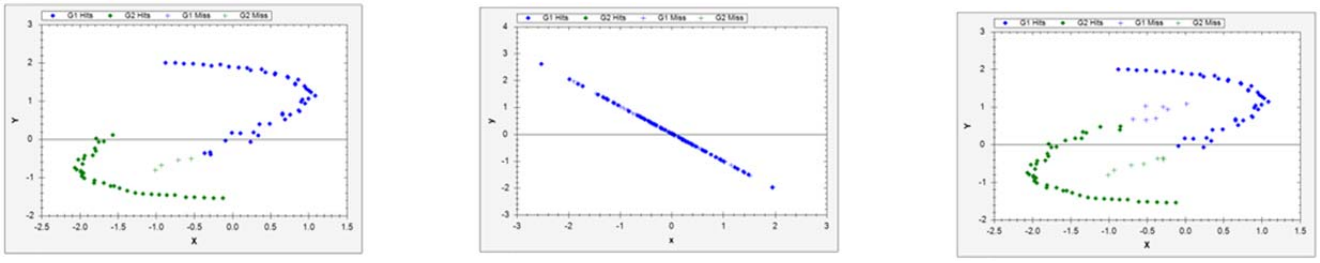
B. Clustering Dataset

For a document corpus with p classes and n documents, we remove the class labels. Then we randomly selected one-third of the documents for training or validation and the remaining for testing. Whereas, the data for training or validation are separate from the data for testing.

In this experiment, we compare the performance of different distance measures. The performance is evaluated by the accuracy, AC, which compares the predicted label of each document with that provided by the document corpus. Figure 1 shows the visualization of performance measures and Table II shows the accuracy results for said three distance measures.

TABLE II. ACCURACY RESULTS

Distance Measure	True Positives	False Negatives	True Negatives	False Positives	Sensitivity	Specificity	Efficiency	Accuracy
Cosine	36	7	43	14	0.8372	0.7543	0.7957	0.79
Euclidean	9	39	52	14	0.1875	0.7878	0.4876	0.5350
SMTP				43	7	43	7	0.86	0.86	0.86	0.86

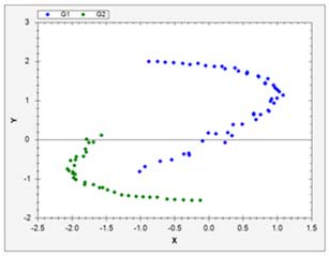


(a) Cosine (b) Euclidean (c) SMTP

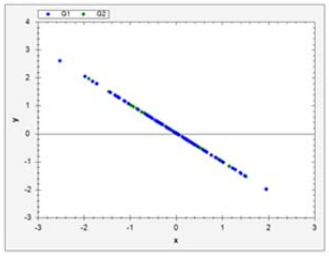
Figure 1. Visualization of Performance Measures

Scatter plot for training data of various distance measures is shown in the Figure 2.

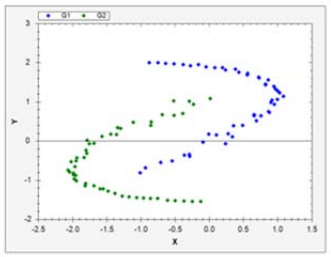
In this experiment we have calculated the values of class probabilities for groups G1 and G2. This class probabilities visualization for various distance measures is shown in the Figure 3.



(a) Cosine



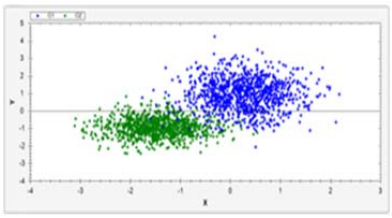
(b) Euclidean



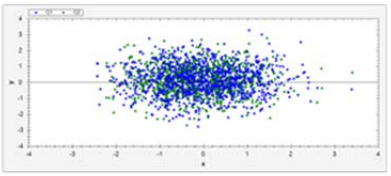
(c) SMTP

Fig. 2. Scatter Plot of Training Data

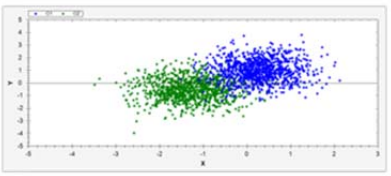
The time required by Naïve Bayes classification and k-NN classification algorithm against variety of training dataset sizes is shown in the Figure 4 & Figure 5 respectively.



(a) Cosine



(b) Euclidean



(c) SMTP

Fig. 3. Class Probabilities of various Distance Measures

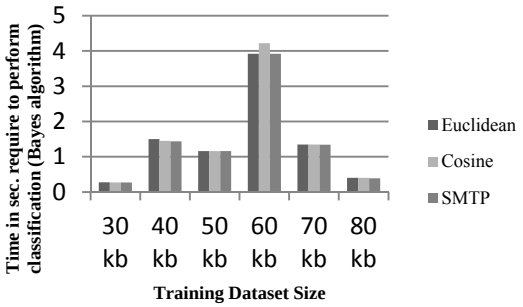


Fig. 4. Time in sec. require to perform classification (Bayes algorithm)

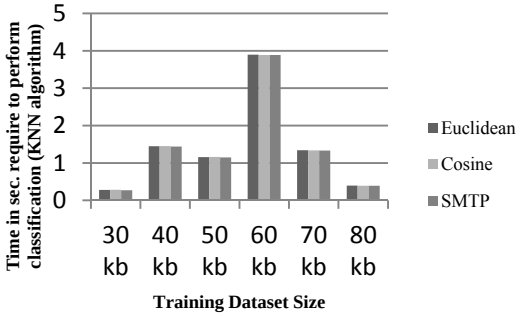


Fig. 5. Time in sec. require to perform classification (KNN algorithm)

V. CONCLUSION

The investigation is done for effectiveness of Euclidean distance; Cosine similarity and SMTP distance measures by applying it in k-NN based classification, Naïve Bayes classification and k-means clustering on real-world data set.

The results have shown that the performance obtained by the SMTP measure is much better than that achieved by other two measures, since accuracy results shown by SMTP are superior as compared to others.

FUTURE WORK

The algorithms and the data sets adopted are intended to be popular and easily accessible for anyone interested in this research area. However, it would be of greater value evaluating the performance of the measures on larger test-beds.

Also, this work mainly focuses on textural features. It would be interesting to investigate the effectiveness and efficiency in the scenarios that involve non-textual features and objects. Besides, as can be seen from the experimental results, the usefulness of a similarity measure could depend on (1) application domains, e.g., text or image, (2) feature formats, e.g., word count or IFIDF, and (3) classification or clustering algorithms. It would be a very interesting topic to examine how certain similarity measures behave in different classification and clustering tasks.

REFERENCES

[1] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, Member, IEEE," A Similarity Measure for Text Classification and Clustering", in *IEEE Transactions On Knowledge and Data Engineering,* Vol. 26, No. 7, July 2014, 1575

[2] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," *IEEE Trans. Knowl. Data Engg.*, vol. 20, no. 9, pp. 1217–1229, Sept. 2008.

[3] Kalaivendhan. K, Sumathi. P, "An Efficient Clustering Method To Find Similarity Between The Documents", in *International Journal of Innovative Research in Computer and Communication Engineering* ,Vol.2, Special Issue 1, March 2014.

[4] B Sindhiya and N Tajunisha1,"CONCEPT AND TERM BASED SIMILARITY MEASURE FOR TEXT CLASSIFICATION AND CLUSTERING", in *Int. J. Engg. Res. & Sci. & Tech.* 2014, Vol. 3, No. 1, February 2014.

[5] K. Sruthi, B. Venkateshwar Reddy, "Document Clustering on Various Similarity Measures", in *International Journal of Advanced Research in Computer Science and Software Engineering,* Volume 3, Issue 8, August 2013 ISSN: 2277 128X.

[6] VenkataGopalaRao S. Bhanu Prasad A., "Space and Cosine Similarity measures for Text Document Clustering", in *International Journal of Engineering Research & Technology (IJERT),*Vol. 2 Issue 2, February- 2013,ISSN: 2278-0181.

[7] P.Sowmya Lakshmi, V.Sushma, T.Manasa, "Different Similarity Measures for Text Classification Using Knn", in*IOSR Journal of Computer Engineering (IOSRJCE)*, Volume 5, Issue 6 (Sep-Oct. 2012), PP 30-36.

[8] Anil Kumar Patidar,JitendraAgrawal,Nishchol Mishra, "Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach", in *International Journal of Computer Applications* ,Volume 40– No.16, February 2012.

[9] Anna Huang , "Similarity Measures for Text Document Clustering", in *New Zealand Computer Science Research Student Conference* 2008, April 2008.

[10] Hung Chim,Xiaotie Deng, "A New Suffix Tree Similarity Measure for Document Clustering" in *ACM 978*-1-59593-654-7/07/0005.

[11] T. Joachims and F. Sebastiani, "Guest editors' introduction to the special issue on automated text categorization," *J. Intell. Inform. Syst.,* vol. 18, no. 2/3, pp. 103–105, 2002.

[12] K. Knight, "Mining online text," *Commun. ACM*, vol. 42, no. 11, pp. 58–61, 1999.

[13] F. Sebastiani, "Machine learning in automated text categorization," *ACM CSUR*, vol. 34, no. 1, pp. 1–47, 2002.

[14] T. Joachims, "A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization," in Proc. 14th *Int. Conf. Mach. Learn*., San Francisco, CA, USA, 1997, pp. 143–151.

[15] H. Kim, P. Howland, and H. Park, "Dimension reduction in text classification with support vector machines," *J. Mach. Learn. Res*., vol. 6, pp. 37–53, Jan. 2005.

[16] G. Salton and M. J. McGill, "Introduction to Modern Retrieval,"*London,* U.K.: McGraw-Hill, 1983.

[17] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining," Boston, MA, USA: Addision-Wesley, 2006.

[18] M. L. Zhang and Z. H. Zhou, "ML-kNN: A lazy learning approach to multi-label learning," *Pattern Recognit.,* vol. 40, no. 7, pp. 2038–2048, 2007.

[19] P. K. Agarwal and C. M. Procopiuc, "Exact and approximation algorithms for clustering," in Proc. 9th *Annu. SODA,* Philadelphia, PA, USA, 1998, pp. 658–667.

[20] W. Aha, "Lazy learning: Special issue editorial," Artif. *Intell.Rev.,* vol. 11, no. 1–5, pp. 7–10, 1997.

[21] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[22] H. Chim and X. Deng, "Effcient phrase-based document similarity for clustering," *IEEE Trans. Knowl. Data Eng.,*vol.20,no.9, pp. 1217–1229, Sept. 2008.

[23] G. Salton, "Automatic Text Processing," Addison-Wesley, New York, 1989.

[24] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," AAAI-2000: *Workshop on Artificial Intelligence for Web Search*, July 2000.

[25] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," Behav. Sci., vol. 12, no. 2, pp. 153–155, 1967.

[26] R. O. Duda, P. E. Hart, D. J. Stork, "Pattern Recognition," New York, NY, USA: Wiley, 2001.

[27] T. W. Schoenharl, G. Madey, "Evaluation of measurement techniques for the validation of agent-based simulations against streaming data," *Proc. ICCS, Kraków, Poland*, 2008.

[28] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," 2nd ed. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA, USA: Elsevier, 2006.

[29] [Online]. Available: http://web.ist.utl.pt/~acardoso/ datasets/

[30] [Online]. Available: http://www.cs.technion.ac.il/ ~ronb/thesis.html

[31] D. D. Lewis, Y. Yang, T. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.,* vol. 5, pp. 361–397, Apr. 2004.

[32] B. Larsen, C. Aone, "Fast and effective text mining using linear-time document clustering," Proceedings of the Fifth *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999.

[33] M. Ester, H.-P.Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proceedings of 2nd International Conference on KDD,* 1996.

[34] B. Larsen, C. Aone, "Fast and e?ective text mining using linear-time document clustering," Proceedings of the Fifth *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999.

[35] R. B. Yates, B. R. Neto, "Modern Information Retrieval," ADDISON-WESLEY, New York, 1999.