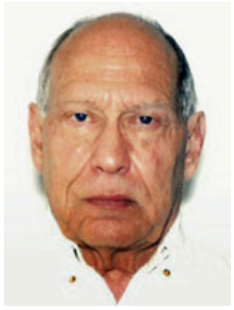
The Use of Learning Techniques to Analyze the Results of a Manual Classification System

Zainab M. AlQenaei* and David E. Monarchi**

*Kuwait University, College of Business Administration, Department of Quantitative Methods & Information Systems, P.O. Box 5486, Safat Kuwait 13055, <zalbader@cba.edu.kw>
** University of Colorado at Boulder, Leeds School of Business, Management and Entrepreneurship Division, <David.E.Monarchi@Colorado.edu>

Zainab M. AlQenaei is an assistant professor of information systems at the College of Business Administration, Kuwait University. Her current research interests include text mining and natural language processing. She received a PhD in business administration (2009) from the University of Colorado, an MBA (2004) from the University of Pittsburgh, and a bachelor's degree in computer engineering (2001) from Kuwait University.

David E. Monarchi is a retired full professor of information systems at the Leeds School of Business, University of Colorado at Boulder. He has published in a variety of peer-reviewed journals. His current research interests are in the areas of text mining and network analysis. He is now an independent consultant working broadly in the area of Business Intelligence.



AlQenaei, Zainab M. and Monarchi, David E. "The Use of Learning Techniques to Analyze the Results of a Manual Classification System." *Knowledge Organization* 43 no. 1: 56-63. 19 references.

Abstract: Classification is the process of assigning objects to pre-defined classes based on observations or characteristics of those objects, and there are many approaches to performing this task. The overall objective of this study is to demonstrate the use of two learning techniques to analyze the results of a manual classification system. Our sample consisted of 1,026 documents, from the ACM Computing Classification System, classified by their authors as belonging to one of the groups of the classification system: "H.3 Information Storage and Retrieval." A singular value decomposition of the documents' weighted term-frequency matrix was used to represent each document in a 50-dimensional vector space. The analysis of the representation using both supervised (decision tree) and unsupervised (clustering) techniques suggests that two pairs of the ACM classes are closely related to each other in the vector space. Class 1 (Content Analysis and Indexing) is closely related to Class 3 (Information Search and Retrieval), and Class 4 (Systems and Software) is closely related to Class 5 (Online Information Services). Further analysis was performed to test the diffusion of the words in the two classes using both cosine and Euclidean distance.

Received: 3 July 2015; Revised 5 October 2015; Accepted 14 November 2015

Keywords: class, classification, classes, correlation, document clustering

1.0 Introduction

Classification is the process of assigning objects to predefined classes based on observations or characteristics of those objects, and there are many approaches to performing this task. Gordon (1999) makes an important point about ways to classify objects. He mentions that a set of objects could be classified in different ways depending upon which characteristic(s) were used to explain the object. Thus, careful thought should be given when selecting the set of variables that will be used to describe the objects. For the purpose of this study, we will be using a 50-dimensional vector from a singular value decomposition (SVD) to represent each document in a vector space model.

Walt and Barnard (2006) studied the relationship between the distribution of data and classifier performance for non-parametric classifiers. The experiments performed on the data show that "predictable factors such as the available amount of training data (relative to the

dimensionality of the feature space), the spatial variability of the effective average distance between data samples, and the type and amount of noise in the data set influence such classifiers to a significant degree." Glanzel and Schubert (2003) proposed a classification system for papers in the sciences, social sciences, arts and humanities. The goal of the researchers was to classify a given document into a defined category using a three-step iterative process. As the authors mention, the results of article classification assisted in determining the disciplinary affiliation of their authors. In other research related to classification schemes, Zins (2007) documented 28 classification schemes in the information science field. The results of the study assisted in further exploring the foundations of this field. The main objective of Janssens et al. (2009, 90) research on classification schemes was to "compare (hybrid) cluster techniques for cognitive mapping with traditional 'intellectual' subject-classification schemes." The authors found that the hybrid clustering techniques applied to a set of journals is superior to other methods used and also allows the improvement of existing classification schemes.

Gordon (1999) explains different aims of classification. First, classification allows the summarization of datasets and also helps to detect relationships and structure within a dataset. This is also a feature of clustering. Gordon (1999) uses the terms classification and clustering interchangeably. Second, if after a classification is performed there still exist objects that could not be assigned to any of the classes, then they could be grouped together with specific properties that define them. Third, if a class contains a group of objects and each object has a specific property, then it would be easier for the researcher to define a name for the group of properties that explain a given class. In this way, new properties could be discovered for certain objects that were not explicit in the object, but because they appeared with other objects in the same group, they would share the same properties. Fourth, classification could allow researchers to frame general hypotheses to account for the observed data in a study.

In 1965, Taulbee and House (1965, 132) presented a paper in the ACM 20th National Conference where they discussed classification in the area of information storage and retrieval. They point to the reasons for classification followed by methods of classification and then cover classification in more depth. At the end of their paper they mention, "the question may be asked how can one evaluate a classification scheme?" The four methods they stated to validate a classification scheme are: 1) recourse to an ultimate criterion; 2) consistency arguments; 3) consensus of opinion: and, 4) effective congruence. Taulbee and House suggest that one method of determining consensus of opinion is to compare two classification schemes. This paper presents a novel approach to analyze the results of a manual classification system.

2.0 Literature review

A vast amount of research has been done in the area of classifying text-based documents. This includes but not limited to manual classification, automatic classification, and various measures of validation to different classification schemes. Aggarwal and Zhai (2012) list several domains in which text classification is used such as news filtering and organization, document organization and retrieval, opinion mining, and email classification and spam filtering. Aggarwal and Zhai (2012) mention two ways in which text-based documents could be represented: as a bag of words or as strings. Baharudin et al. (2010) discussed techniques and methodologies used in text documents classification. Baharudin et al. (2010, 16) mention that "more works are required for the performance improvement and accuracy of the documents classification process."

In terms of evaluation methods, Aggarwal and Zhai (2012, 209) explain different evaluation methods for text classification such as bagging, stacking, and boosting: "Meta-algorithms play an important role in classification strategies because of their ability to enhance the accuracy of existing classification algorithms by combining them, or making a general change in the different algorithms to achieve a specific goal." Pong et al. (2008, 219) test two supervised machine learning algorithms for automatic document classification and state that "such a complex categorization scheme, developed for manual document classification, may not be suitable for automatic document classification." Another study (Desale and Kumbhar 2013) suggests that the use of automated classification scheme using natural language and artificial intelligence. Roitblat et al. (2010, 73) compare two categorization processes "with the more traditional process of having people, usually lawyers, read and categorize each document. This study uses agreement to assess the level of reliability of the human and computer processes." Authors conclude that the performance of the two computer systems used to categorize text-based documents was at least as accurate of that of human review. Other research (e.g. Al-Ghuribi and Alshomrani 2014, Luo and Li 2014, and Pong et al. 2008), uses well-known measures for automatic classifiers (precision, recall, and F-1 measures) to compare the results of different classifiers. In a different approach to improve a classification scheme, Janssens et al. (2009) explore the possibility of using the results of a cluster analysis.

Some studies propose a new hybrid method of classifying text documents (e.g. Ur-Rahman and Harding 2012), others compare hybrid clustering techniques with tradi-

tional subject-classification schemes (Janssens et al. 2009). In their research, Ur-Rahman and Harding (2012) first classified documents manually by domain experts then used the term-frequency (TF) representation of textual documents. Glanzel and Schubert (2003, 364) state a crucial point: "All papers published in journals not assignable to 'well-defined' subject categories have to be assigned individually, i.e., paper by paper." This raises the importance of having a well-defined technique to analyze the results of a classification scheme, which is the purpose of this research.

Other research focuses on using natural language processing (NLP) with statistics in the context of text categorization and classification. For example, Jacobs' research focuses on (1992, 78) "combining statistics and NLP in a knowledge-based categorization system, using statistics as a way of augmenting hand-coded knowledge." Our research is similar to Jacobs (1992), Wiebe et al. (1999), and many others in the sense that NLP is used with multivariate statistics. However, our research focuses on using a vector space model and multivariate statistics to assess an existing manual text classification system. In the next section we will briefly give an example of the process by analyzing a section of the ACM Computing Classification System using both supervised (decision tree) and unsupervised (clustering) learning techniques.

3.0 Data

The ACM provides a link on "How to Classify Works Using ACM's Computing Classification System" for authors to follow and decide what the most appropriate category for their paper is. The link for the guidelines is: http://www.acm.org/class/how_to_use.html. Basically, the papers in the ACM classification systems are manually classified, and we present a method to analyze the classification.

The ACM Computing Classification System (1998) can be found at http://www.acm.org/class/1998/TOP.html. We have collected data on only class H.3 Information Storage and Retrieval. The following are subclasses in H.3:

- H.3.0 General
- H.3.1 Content Analysis and Indexing - *Class 1*
- H.3.2 Information Storage – *Class 2*
- H.3.3 Information Search and Retrieval - *Class 3*
- H.3.4 Systems and Software – *Class 4*
- H.3.5 Online Information Services – *Class 5*
- H.3.6 Library Automation – *Class 6*
- H.3.7 Digital Libraries – *Class 7*
- H.3.m Miscellaneous

The sub-classifications not used after 1998 were not considered in the analysis, nor were the general and miscellaneous levels. Only the seven classes H.3.1—H.3.7 in the list above were used. The next step was to search in the ACM database in the attribute "classification" for the seven classes above as the primary classification. A stratified random sample was taken across the seven classes. A total of 1,026 documents (abstracts) were collected and used in the analysis described below.

The ACM has published a newer system in 2012, and according to ACM's website: "The old scheme has been mapped to the new, and both the 1998 and 2012 terms are available on Citation Pages of all indexed articles in the ACM Digital Library." For the purpose of this research using any scheme would fulfill the objectives of the analysis.

4.0 Methodology

4.1 Document representation

Document representation includes the following steps: preprocessing, obtaining the term-frequency matrix, transformation, and decomposition. Preprocessing the documents involves removing stopwords, stemming the remaining words, and identifying parts of speech to be used in the analysis (e.g. ignoring adverbs). Stopwords include general words such as the, an, is, at, etc. They also include common domain-specific words (e.g., information). Lemmatization is the process of reducing a word to its original root. For example, the root of the word "processing" is "process." In this manner the multiple forms of a word (its morphology) will be reduced to the same root. We used the SAS Enterprise Miner 5.2 Text Miner node to determine the parts of speech and to lemmatize the terms. The parts of speech that were kept as part of this step were the nouns and verbs. The rest were ignored.

Next, the term by document frequency (TF) matrix was constructed. This matrix (X) has the terms as rows and the documents as the columns (i.e., a 844 by 1026 matrix). The cells are the count (frequency) of a term in a document. The counts are weighted using the log-entropy weighting method (log as the local weighting and entropy as the global weighting for each term). The local weighting increases or decreases the importance of a term within a document while the global weighing does the same thing but across the whole corpus. The formulas used for the log-entropy weighting are (Hare and Lewis 2005):

$$L(i, j) = \log (tf_{ij} + 1) \qquad \text{Local weighting}$$

$$\sum^{N} [(\left(\frac{tf_{ij}}{gf_i} \right)] \log [(\left(\frac{tf_{ij}}{gf_i} \right)]))$$

$$G(i) = 1— / \log N \qquad \text{Global weighting}$$

where tf_{ij} is the frequency of term i in document j, gf_i is the frequency of the term i in the entire corpus, and N is the number of documents (which in this case is 1,026).

Another common transformation applied to the term-frequency matrix is normalization, which converts the document vectors to a unit length thereby compensating for the varying number of words in the documents. The result of weighting and normalizing X is the matrix A. Finally, A was decomposed into three matrices using singular value decomposition (SVD). This is described in the next section.

4.2 Matrix decomposition

Singular value decomposition (SVD) (Golub and Van Loan 1996) is a mathematical technique that decomposes a rectangular matrix into a linear combination of three matrices. The decomposition of A results in the following three matrices: U, S, and V.

$$A_{m \times n} = U_{m \times r} S_{r \times r} V^T_{r \times n} \approx U_{m \times k} S_{k \times k} V^T_{k \times n} = \hat{A}_{m \times n}$$

Where:

U: left singular vectors corresponding to the terms
V: right singular vectors corresponding to the documents
S: singular values
r: rank of A, which is ≤ (m, n)
k: number of dimensions retained, k ≤ r
Â: approximation of A using k dimensions

In this representation, each term is represented by a row in U. Similarly, each document is represented by a row in V.

The critical point here is the issue of dimension reduction: that is, the reduction of the dimensionally of the vector space from r to k dimensions. The choice of k has been mainly a matter of judgment by the researchers. As Deerwester et al. (1990, 398) state, "we want a value of k that is large enough to fit all the real structure in the data, but small enough so that we do not also fit the sampling error or unimportant details. The proper way to make such choices is an open issue in the factor analytic literature." Deerwester et al. (1990) used 50-100 dimensions in their first study. The SVD of text data has been used in many applications with k typically between 50 and 500. In this study we retained 50 dimensions and used them as inputs to the decision tree and the clustering analyses discussed below.

In order to analyze the classification of ACM's H.3 section, a supervised learning technique (decision tree) and an unsupervised learning technique (clustering) were used. We describe both in the following sections.

4.3 Decision tree

The 1,026 50-dimensional vectors representing the documents were analyzed using a decision tree algorithm. A stratified sample of 70% of the vectors was used as the training dataset, and 30% was used as the validation dataset. The validation data set was used for monitoring and tuning the decision tree model to improve its generalization. The target variable was the ACM "class" of each document, ranging from 1 to 7, the H.3 subclasses. Since it is a nominal target variable, the default splitting criterion used in SAS E-Miner is the Chi-square test. According to SAS, "the data is partitioned according to the best split. The process repeats in each leaf until there are no more allowed splits" The threshold is the minimum acceptable p-value. We used the default value of 0.2.

4.4 Clustering

The same 1,026 50-dimensional vectors representing the documents were clustered using the Expectation Maximization Method in SAS. The algorithm produces exactly *k* different clusters; the default value of k in SAS is 10, which was used in this research. The algorithm computes probabilities of cluster memberships based on one or more probability distributions. The clustering is modeled using a Gaussian probability distribution.

5.0 Results

5.1 Centroids of members of each class

Each document is represented by a 50-dimensional vector in the space created by the SVD. The first step taken was to find how close the centroids of the documents in each of the seven classes were. Both the cosine similarity measure and the Euclidean distance were used. The results are shown in Table 1. The centroids of Class 1 and Class 3 are closer to each other than to any others, and the same applies to Class 4 and Class 5. An important observation is that both the cosines and the Euclidean distance measures produce the same results.

5.2 Output of the decision tree

The result of the decision tree analysis of the validation dataset is shown in Figure 1. Figure 1 shows the distribution of classes 1, 3, 4, and 5 among the 11 leaf nodes. Leaf node 4 contains the highest percentage of classes 1 and 3. Leaf node 5 contains the highest percentage of classes 4 and 5.

	Class 1		Class 2		Class 3		Class 4		Class 5		Class 6		Class 7	
	Cosine	ED	Cosine	ED	Cosine	ED	Cosine	ED	Cosine	ED	Cosine	ED	Cosine	ED
Class 1	1.000	.000	.856	.584	.978	.217	.901	.461	.870	.532	.853	.579	.847	.586
Class 2	.856	.584	1.000	.000	.833	.626	.883	.526	.836	.622	.832	.639	.811	.673
Class 3	.978	.217	.833	.626	1.000	.000	.926	.399	.896	.474	.854	.576	.855	.568
Class 4	.901	.461	.883	.526	.926	.399	1.000	.000	.960	.294	.861	.561	.875	.527
Class 5	.870	.532	.836	.622	.896	.474	.960	.294	1.000	.000	.851	.583	.886	.505
Class 6	.853	.579	.832	.639	.854	.576	.861	.561	.851	.583	1.000	.000	.956	.320
Class 7	.847	.586	.811	.673	.855	.568	.875	.527	.886	.505	.956	.320	1.000	.000

Table 1. Similarity (cosines) and dissimilarity (Euclidean distance) matrix

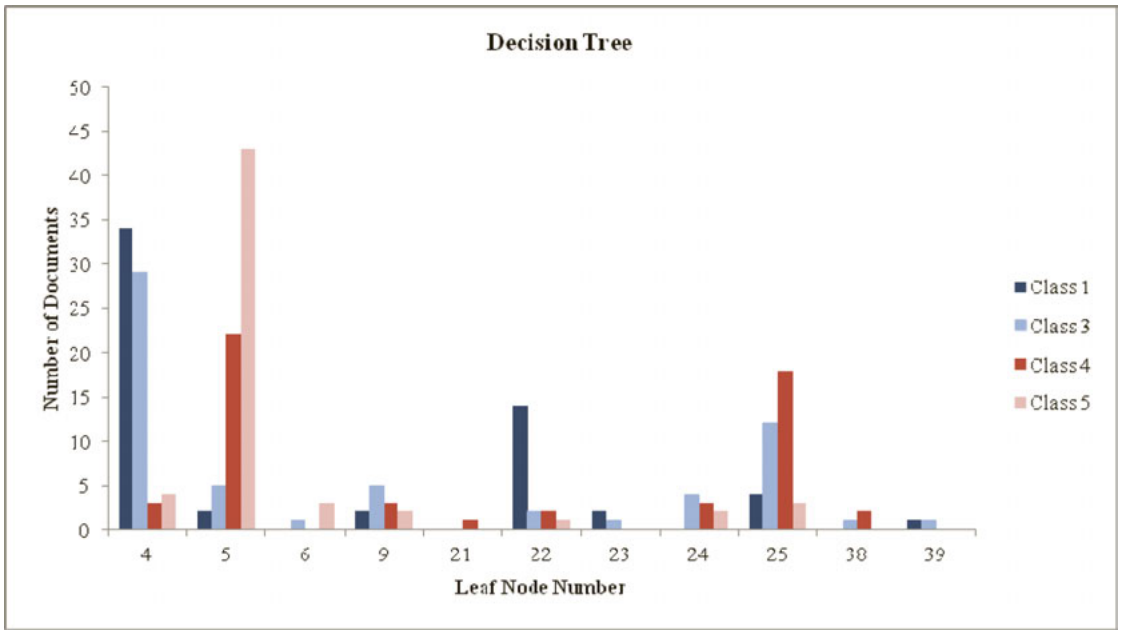


Figure 1. Decision tree

5.3 Output of the clustering

The distribution of the 1,026 abstracts among the 13 clusters is shown in Figure 2. Again, classes 1 and 3 and classes 4 and 5 are paired together. The distribution of the documents across the 13 clusters shows that the highest percentage of classes 1 and 3 both occur in clusters 2 and 5. The highest percentage of classes 4 and 5 both occur in clusters 3 and 5.

5.4 Correlation coefficient

The analysis of the correlation coefficients show that from the distribution of the components of the classes across the leaf nodes in the decision tree (Table 2), Class 1 and Class 3 are significantly correlated (r=0.865, p=0.001), Class 4 and Class 5 (r=0.774, p=0.005). In the clustering analysis the results again show that from the distribution of the components of the classes across the clusters (Table 3), Class 1 and Class 3 are significantly correlated (r=0.857, p<0.0001), as are Class 4 and Class 5 (r=0.901, p=<0.0001).

5.5 Chi-square to test the distribution of classes

We have demonstrated that classes 1 and 3 and classes 4 and 5 are strongly related in terms of their distributions among both the decision tree leaf nodes and the clusters. Next we will test the existence of a leaf-node or a cluster effect using a Chi-square goodness-of-fit test. The null hypothesis is that the classes are distributed evenly among the leaf nodes and clusters. The results indicate that there is indeed both a leaf-node effect and a cluster effect. In the decision tree, the reported $\chi^2(6, N=59) = 104.847$, $p = 0.000$ for class 1, the reported $\chi^2(9, N=61) = 112.607$, $p = 0.000$ for class 3, the reported $\chi^2(7, N=54) = 71.037$, $p = 0.000$ for class 4, the reported $\chi^2(6, N=58) = 170.345$, $p = 0.000$ for class 5, and the reported $\chi^2(8, N=55) = 85.236$, $p = 0.000$ for class 7. In the clustering, the reported $\chi^2(13, N=194) = 135.072$, $p = 0.000$ for

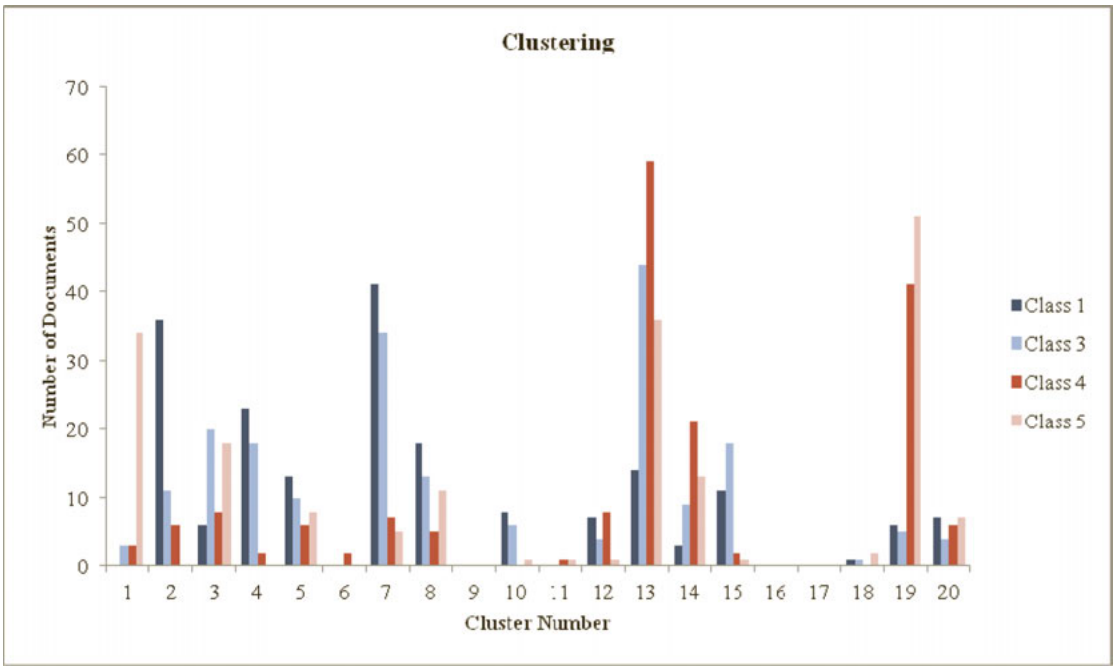


Figure 2. Cluster analyses

		Class_1	Class_2	Class_3	Class_4	Class_5	Class_6	Class_7
Class_1	Pearson Correlation	1	.020	.865**	-.048	-.045	.185	-.042
	Sig. (2-tailed)		.954	.001	.889	.894	.586	.903
Class_2	Pearson Correlation	.020	1	.113	.684*	.840**	.581	.553
	Sig. (2-tailed)	.954		.741	.020	.001	.061	.078
Class_3	Pearson Correlation	.865**	.113	1	.227	.064	.075	.046
	Sig. (2-tailed)	.001	.741		.502	.852	.826	.892
Class_4	Pearson Correlation	-.048	.684*	.227	1	.774**	.238	.538
	Sig. (2-tailed)	.889	.020	.502		.005	.481	.088
Class_5	Pearson Correlation	-.045	.840**	.064	.774**	1	.590	.779**
	Sig. (2-tailed)	.894	.001	.852	.005		.056	.005
Class_6	Pearson Correlation	.185	.581	.075	.238	.590	1	.843**
	Sig. (2-tailed)	.586	.061	.826	.481	.056		.001
Class_7	Pearson Correlation	-.042	.553	.046	.538	.779**	.843**	1
	Sig. (2-tailed)	.903	.078	.892	.088	.005	.001	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 2. Decision tree correlation coefficients

class 1, the reported $\chi^2(10, N=65) = 122.338$, $p = 0.000$ for class 2, the reported $\chi^2(14, N=200) = 153.55$, $p = 0.000$ for class 3,the reported $\chi^2(14, N=177) = 325.996$, $p = 0.000$ for class 4, the reported $\chi^2(13, N=189) = 241.593$, $p = 0.000$ for class 5, the reported $\chi^2(6, N=21) = 26.67$, $p = 0.000$ for class 6, and the reported $\chi^2(14, N=180) = 491.67$, $p = 0.000$ for class 7. There were no leaf-node effects for the other classes.

6.0 Conclusions and future work

In this paper, we analyzed the results of a manual classification system (a section of the ACM Computing Classification System) using both supervised (decision tree) and unsupervised (clustering) learning techniques. The data consisted of 1,026 documents self-classified by each document's author(s) as belonging to one of the seven clas-

		Class_1	Class_2	Class_3	Class_4	Class_5	Class_6	Class_7
Class_1	Pearson Correlation	1	.102	.639	.089	-.089	-.081	-.078
	Sig. (2-tailed)		.670	.002	.710	.709	.742	.744
Class_2	Pearson Correlation	.102	1	.192	.292	.156	.278	.086
	Sig. (2-tailed)	.670		.416	.212	.511	.249	.719
Class_3	Pearson Correlation	.639	.192	1	.565	.296	.022	.152
	Sig. (2-tailed)	.002	.416		.009	.205	.929	.523
Class_4	Pearson Correlation	.089	.292	.565	1	.775	.173	.296
	Sig. (2-tailed)	.710	.212	.009		.000	.479	.205
Class_5	Pearson Correlation	-.089	.156	.296	.775	1	.142	.229
	Sig. (2-tailed)	.709	.511	.205	.000		.561	.332
Class_6	Pearson Correlation	-.081	.278	.022	.173	.142	1	.967
	Sig. (2-tailed)	.742	.249	.929	.479	.561		.000
Class_7	Pearson Correlation	-.078	.086	.152	.296	.229	.967	1
	Sig. (2-tailed)	.744	.719	.523	.205	.332	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3. Clustering correlation coefficients

ses of the Classification System: H.3 Information Storage and Retrieval. Our analyses are based on the singular value decomposition (SVD) of the 1,026 documents' term-frequency (TF) matrix (i.e., a vector space model).

We have evidence to state that two pairs of the classes are closely related to each other in the vector space of the SVD. Class 1 is closely related to Class 3, and Class 4 is closely related to Class 5. We examined the physical proximity of the classes using the cosine similarity measure and the Euclidean distance to determine how close the centroids of each of the seven classes were. Using both measurements, the centroids of Class 1 and Class 3 are closer to each other than to any others, and the same applies to Class 4 and Class 5. To examine the distribution of the abstracts in the classes, we used the correlation coefficient to observe that classes 1 and 3 and classes 4 and 5 are strongly related in terms of their distributions among both the decision tree leaf nodes and the clusters. Also, we have demonstrated that there exist both a leaf node effect and a cluster effect by using a Chi-square test. Since the two pairs of classes are behaving similarly across the nodes we could say that they are addressing closely related topics.

The results show some degree of overlap among clusters. Class 1 and Class 3 seem to fit together. However, this is less true for classes 4 and 5. One explanation might be that the keywords supplied to the authors for them to assign their paper to Class 4 or Class 5 may not have crisply differentiated between the classes and consequently confused the authors in their selection of a class for their paper. Our analysis of the abstracts and the cluster overlap would then reflect that fuzziness. Alternatively, a new classification could be merging with bits of both classes creating a new class in the classification scheme.

Regardless of the keywords that are contained in the papers (abstracts) in both classes, we have reason to believe that classes 1 and 3, and classes 4 and 5 semantically go together. That is, both pairs are discussing the same topics. Future work could be performed to further analyze the subclasses and examine the keywords that appear in both pairs of classes. Furthermore, as there exists a variety of clustering algorithms, one could examine the results suggested by a different algorithm.

References

Aggarwal, Charu C. and ChengXiang Zhai. 2012. "A Survey of Text Classification Algorithms." In *Mining Text Data,* edited by Charu C. Aggarwal and ChengXiang Zhai. New York: Springer, 163-222.

Al-Ghuribi, Sumaia Mohammed and Saleh Alshomrani. 2014. "Bi-Languages Mining Algorithm for Classifying Text Documents (BiLTc)." *International Journal of Academic Research* 6, no. 5: 16-25.

Baharudin, Baharum, Lam Hong Lee and Khairullah Khan. 2010. "A Review of Machine Learning Algorithms for Text-Documents Classification." *Journal of Advances in Information Technology* 1, no. 1: 4-20.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas and Richard A. Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41: 391-407.

Desale, Sanjay K. and Rajendra M. Kumbhar. 2013. "Research on Automatic Classification of Documents in Library Environment: A Literature Review." *Knowledge Organization* 40: 295-304.

Glänzel, Wolfgang and András Schubert. 2003. "A New Classification Scheme of Science Fields and Subfields Designed for Scientometric Evaluation Purposes." *Scientometrics* 56: 357-67.

Golub, Gene H. and Charles F. Van Loan. 1996. *Matrix Computations*. 3rd ed. Baltimore: Johns Hopkins University Press.

Gordon, A. D. 1999. *Classification*. 2nd ed. Boca Raton : Chapman & Hall/CRC.

Hare, Jonathon S. and Paul H. Lewis. 2005. "On Image Retrieval Using Salient Regions with Vector-Spaces and Latent Semantics." In *Image and Video Retrieval: 4th International Conference, CIVR 2005, Singapore, July 2005,* edited by Erwin M Bakker, Lekha Chaisorn, Tat-Seng Chua, Wee-Kheng Leow, Michael S Lew and Wei-Ying Ma. Lecture Notes in Computer Science, 3568. Berlin; Heidelberg: Springer, 540-9.

Jacobs, Paul S. 1992. "Joining Statistics with NLP for Text Categorization." *Proceedings of the Third Conference on Applied Natural Language Processing. March 31-April 3, 1992 Trento, Italy*, edited by Madeleine Bates and Oliviero Stock. Stroudsburg, PA: Association for Computational Linguistics, 178-85.

Janssens, Frizo, Lin Zhang, Bart De Moor and Wolfgang Glänzel. 2009. "Hybrid Clustering for Validation and Improvement of Subject-Classification Schemes." *Information Processing & Management* 45: 683-702.

Luo, Le and Li Li. 2014. "Defining and Evaluating Classification Algorithm for High-Dimensional Data Based on Latent Topics." *PloS One* 9, no.1: e82119.

Pong, Joanna Yi-Hang, Ron Chi-Wai Kwok, Raymond Yiu-Keung Lau, Jin-Xing Hao, and Percy Ching-Chi Wong. 2008. "A comparative Study of Two Automatic Document Classification Methods in a Library Setting." *Journal of Information Science* 34: 213-30.

Roitblat, Herbert L., Anne Kershaw and Patrick Oot. 2010. "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review." *Journal of the American Society for Information Science and Technology* 61: 70-80.

Taulbee, Orrin E. and R. W. House. 1965. "Invited Papers—1: Classification in Information Storage and Retrieval." In *Proceedings of the 1965 20th national conference, August 24-26, 1965 Cleveland, Ohio,* edited by Lewis Winner,. New York, N.Y.: ACM, 119-37

Ur-Rahman, Nadeem and Jennifer A. Harding. 2012. "Textual Data Mining for Industrial Knowledge Management and Text Classification: A Business Oriented Approach." *Expert Systems with Applications* 39: 4729-39.

Walt, Christiaan van der and Etienne Barnard, 2006. "Data Characteristics that Determine Classifier Performance." *Proceedings of the 17th Annual Symposium of the Pattern Recognition Association of South Africa, November 29-December 1 2006, Parys, South Africa.* SAIEE Africa Research Journal, vol. 98, no. 3. Gardenview [South Africa]: SAIEE Publications, 87-93.

Wiebe, Janyce M., Rebecca F. Bruce and Thomas P. O'-Hara. 1999. "Development and Use of a Gold-Standard Data Set for Subjectivity Classifications." In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 20-26 June 1999, University of Maryland, College Park, Maryland.* edited by Robert Dale and Ken Church. [New Brunswick, N.J.]: Association for Computational Linguistics, 246-53.

Zins, Chaim. 2007. "Classification Schemes of Information Science: Twenty-Eight Scholars Map the Field." *Journal of the American Society for Information Science & Technology* 58: 645-72.