EVALUATING AND COMPARING TEXT CLUSTERING RESULTS

Louis Massey
Royal Military College
PO Box 17000 Stn Forces
Kingston, ON, Canada, K7K 7B4
MASSEY-L@rmc.ca

ABSTRACT
Text clustering is a useful and inexpensive way to organize vast text repositories into meaningful topics categories. However, there is little consensus on which clustering techniques work best and in what circumstances because researchers do not use the same evaluation methodologies and document collections. Furthermore, text clustering offers a low cost alternative to supervised classification, which relies on expensive and difficult to handcraft labeled training data. However, there is no means to compare both approaches and decide which one would be best in a particular situation. In this paper, we propose and experiment with a framework that allows one to effectively compare text clustering results among themselves and with supervised text categorization.

KEYWORDS
Text clustering, text categorization, evaluation.

1. Introduction

Clustering is the operation by which similar objects are grouped together in an unsupervised manner [1]. When clustering textual data, one is mining for relationships among documents. Clustering outputs sets of documents with similar content, the clusters thus representing topics. In this paper, we consider one of the many applications of clustering in the fields of information retrieval and text mining, namely clustering that aims at self-organizing a textual document collection. This application of clustering can be seen as a form of classification by topics, hence making it the unsupervised counterpart of text categorization [2].

The operating concept of a text clustering system is that instead of searching by keywords or exploring the whole collection of documents, a user can browse the clusters to identify and retrieve relevant documents [3]. As such, clustering provides a summarized view of the information space by grouping documents by topics. A representative but by no means exhaustive list of work on text clustering includes [4, 5, 6].

The main purpose of text clustering in corporations or governments is to organize large document collections that change rapidly and that are impossible to organize manually. This is true of the Internet, but also of organization's intranets, of document management systems and even of employees hard disks. From an information and knowledge management point of view, all these are in fact repositories of documents that contain a very large amount of essential corporate knowledge. Without adequate access to that knowledge, rework and other inefficiencies are inevitable and may negatively affect the competitive position of a corporation or efficiency of service providing organizations in government.

Clustering is often the only viable solution to organize large, dynamic text collections by topics. Indeed, even supervised text categorization, although shown to have achieved high accuracy [7], is sometimes impossible to apply because of the high cost and difficulties associated with the handcrafting of large sets of labeled examples to train the system, which partly defeats the purpose of automation. Furthermore, due to the often-changing nature of the document collection, regular, time consuming re-training is needed [8, 9]. The advantage of clustering is thus realized when a training set and class definitions are unavailable, or when creating them is either cost prohibitive due to the collection shear size or unrealistic due to the rapidly changing nature of the collection.

2. Problem Statement and Contributions

Although text clustering can be seen as an alternative to supervised text categorization, the question remains of how to determine if the resulting clusters are of sufficient quality compared to what can be achieved with supervised techniques to be useful in a real-life application. In short, one needs a means of evaluating clusters quality and comparing it with the quality of classes generated by supervised algorithms. This is a critical requirement for any organization contemplating the implementation of a document classification or clustering system.

Furthermore, another problem with text clustering is that there is little knowledge of what works best and in what circumstances. A lack of a commonly accepted experimental methodology is primarily responsible for this situation. Indeed, studies published on text clustering use various data sets and evaluation methods to compare algorithms. Then, how can one claim that one algorithm works well or better than another one? The current situation has impeded the scientific development of the important research field that text clustering is.

Our contribution is to propose an experimental and evaluation framework that allows comparison of various clustering results among themselves and also with supervised classification results. Our proposal is inspired by the tremendous scientific developments achieved over the last decade with supervised text categorization. These developments were largely due to a widely accepted experimental methodology, particularly with respect to benchmark text collections and well established evaluation methodologies. The fact that we use an experimental approach similar to supervised text classification allows for methodological unification with that field, another major contribution and advantage given the recent interest in co-training with unlabeled data due to the high cost of training set labeling

3. Proposed Experimental Methodology

3.1 Benchmark data

We propose the use of the text categorization benchmark Reuter-21578 Distribution 1.0[1] corpus, and in particular of the so called "ModApté" split (hereafter "Reuter"). It is essential that the instructions accompanying the data set be followed precisely to generate exactly the correct set of documents to ensure all research is conducted with the same documents set. The Reuter data has been used extensively in supervised text categorization [10, 11, 12, 13, 7] and has contributed, through standardized experimental work, to highlight good practices in that field. Reuter is known to be challenging because of skewed class distribution, multiple overlapping categories, noisiness and real-life origin (Reuter newswires during the year 1987, in chronological order). Hence, it can be artificially streamed [14] since it is time stamped, which is practical for incremental and on-line clustering experiments. Reuter provides the following pre-established document sets for supervised classification: training set (9,603 documents), test set (3,299 documents) and discarded (8,676 documents). For clustering, the training set is not required; so the only data that should be involved in the clustering itself is the test set. However, the training and discarded sets could be used to accumulate word statistics to perform feature selection. What is essential is that the exact same test documents set be used for all text clustering experiments. This ensures results are comparable between studies. 93 topics are pre-defined in the data. This value is often required by clustering algorithms.

In each experiment, a second data set should also be tested. This second data sets is necessary to eliminate any out of the ordinary good (or bad) performance that may be caused by "compatibility" (or lack thereof) between algorithm and data. To this effect, we propose the use of the HD-49 data set. HD49 is a subset of the larger OHSUMED corpus [15]. Again, it is a collection popular in text categorization [16, 17, 18]. HD49 is recognized as more difficult that Reuter and has consistently given lower quality classification in the supervised case. HD49 is a collection of 3653 abstracts in the medical domain (specifically on Heart Diseases, hence HD) from 1987 to 1991. The test set is comprised of documents from year 1991, and once gain only the test set must be used for clustering. 49 topics are predefined. HD49 is also valuable for experimental work on hierarchical clustering since the original topics are organized hierarchically.

A future issue that will need to be considered is scaling up to larger text collections. The two corpus suggested here contain a mere 3000 documents or so. Most real life applications work with collections several orders of magnitude larger. A third benchmark text collection containing, for example, over one million documents will most likely also be required for realistic evaluations.

3.2 Common evaluation method

The choice of a clustering quality evaluation metric usually depends on the application [19]. In the context of text clustering, several options are available. One is to compare clustering results to an existing solution prepared manually by professional indexers. This approach is known as *external cluster validity* [1]. We can also measure quality as how well clusters are separated and how compact they are. This is *internal cluster validity* [1]. We can also evaluate clustering with users, given a certain task to perform. Since different documents organizations are possible, users studies may seem like the option of choice to validate these other ways to organize the documents. However, user studies are costly and may also be subjective. Because of these difficulties, we consider this approach to be impractical for most researchers.

Internal validity quality measures consider structural aspects of clusters such as their degree of separation and compactness. This gives an idea of quality that may not correspond to the real-life experience and perception of users. Since we are concerned here with the application of clustering in the context of organizing text data to facilitate finding information, this disconnection between quality and actual usefulness is inappropriate. Should an application require evaluation in an exploratory or

[1] Available from http://www.daviddlewis.com/resources/testcollections/reuters21578/

knowledge discovery context where a known solution does not exist, then internal validity is very applicable. However, in our case we do have existing handcrafted solutions available with our two benchmark text collections, so they can be used to evaluate clustering quality in the external validity framework. Thus, we actually evaluate how well the specified document structure is *recovered* by clustering. This appears to be reasonable since the solution crafted by professional indexers can be assumed to closely match specified customer requirements. Although other possible solutions may exist, one can assume that the one handcrafted by human classifiers is a useful one for a general information access task and for most potential users, and therefore should meet our immediate objective of quality evaluation.

We specifically propose to use the F1 measure [20] to evaluate clustering quality. This measure is widely used in supervised text categorization [2], but also in text clustering [4, 8, 21, 22]. As for most clustering validity measures, this one has its strength and weaknesses. F1 provides a good balance between precision and recall, which is excellent in the context of information retrieval. Furthermore, having been used extensively in text categorization, the application of F1 to evaluate clustering quality on the exact same data sets makes comparison with published text categorization results such as [7] possible, something we have accomplished successfully in previous work [8]. Our objective in performing such comparison is that if text clustering is to become a useful information retrieval and text mining tool, its performance must be clearly established. Since supervised categorization has been shown to achieve quality comparable to human classifiers [2, 23], it makes sense to compare the supervised and unsupervised approaches. Doing so, we determine how well clustering, a very low cost approach, does compared to the more costly and human intervention intensive supervised categorization. It is important however to compare to the best available results in text categorization to avoid misleading results. From a practical standpoint, the comparison could help management make a rational decision between clustering and text categorization based on a comparison of their respective cost and F1 quality values.

This is an innovative and potentially extremely useful aspect of our experimental approach: text categorization F1 quality results are used as an upper bound for cluster quality, since learning in a supervised framework with labeled data should provide the best possible automated text classification. Thus, clustering can be expected to eventually approach this level of quality but not exceed it since it relies solely on the data itself. This way of evaluating clustering quality allows one to clearly establish the level of quality obtained by a clustering algorithm as a percentage of the upper bound quality. In the end, we get a very clear picture of the quality of clustering by comparing all algorithms on a common scale that also unifies clustering with supervised categorization.

4. How to compute F1?

The F1 quality is computed as follows:

$$F_1 = 2pr/(p+r)$$

where p is the precision and r the recall. A value of 1 indicates maximal quality and 0 worst quality. Precision and recall are defined as $p = a/(a + b)$ and $r = a/(a + c)$ where a is the number of true positives, i.e. the total number of documents found together in the provided solution set and that are indeed clustered together by the clustering algorithm; b is the number of false positives, i.e. the number of documents not expected to be found together but that are nevertheless clustered together; and c is the number of false negatives, i.e. the number of documents expected to be grouped together but that are not clustered together by the clustering algorithm.

To compute a, b and c, one needs a handcrafted solution set $S = \{S_j \mid j = 1, 2, \ldots, M^S\}$, where $M^S = |S|$ is the number of topics the professional indexers found in the text collection. Each topic S_j is in turn a set of documents that "belong" to that topic. The output of a text categorization system is a set of classes $C = \{C_i \mid i = 1, 2, \ldots, M^S\}$. Class C_i corresponds to topic S_i. Similarly, the output of a text clustering system is a set of clusters $C = \{C_i \mid i = 1, 2, \ldots, M\}$, where M is the number of clusters found. M^S is usually unknown in the context of clustering, and therefore M may not equal M^S. Further, with clustering there is no guarantee that C_i corresponds to topic S_i. Otherwise, the output of a text clustering algorithm is undistinguishable from the output of a supervised text classification algorithm, that is a set of document sets, which is why we denote both by C. In the case of supervised text categorization, establishing the values a, b and c is straightforward since we know that for each topic S_j the corresponding class will be C_j. Therefore, the only pairs topics-classes (S_j , C_i) that need consideration are those for which i=j. We then have for each topic j:

$a_j = |C_j \cap S_j|$ (number of documents both in C_j and S_i)
$b_j = |C_j| - a_j$
$c_j = |S_j| - a_j$

These values are then assembled into a global F1 value by either macro-average or micro-average [2].

The manner in which a, b and c are computed in text categorization cannot be applied to clustering since we do not a priori know which cluster corresponds to which topic of the handcrafted solution. In other words, one cannot only consider the pairs topics-clusters (S_j , C_i) for which i=j. The approach generally used to compute a, b

and *c* (and hence F1) in text clustering [4, 21, 22]) is to take the best cluster i* (the one with the highest F1 score F_1^{i*}) for each topic j as the cluster matching that topic and perform a weighted average of these best F1 values:

$$F_1 = \frac{\sum_{j=1}^{M^S} |S_j| F_1^{i*}}{\sum_{j=1}^{M^S} |S_j|}$$

This way on computing F1 looks similar in form to macro-averaged F1 for text categorization, but because of the weighting by each topic size $|S_j|$ it should behave rather like micro-average. By considering only the best topic-cluster matches, many false negatives and false positives are not accounted for in the F1 calculation. This may unfairly inflate the quality computed compared to text clustering. A possible solution to this apparent problem is to compute *a*, *b* and *c* utilizing the pair-wise counting procedure common to traditional external cluster validity measures [24, 25], but combine these values following the F1 formulae.

We verified these assertions experimentally by comparing micro- and macro-averaged text categorization F1 with F1 based on best cluster-topic matches (F1best) and F1 computed with pair-wise *a*, *b* and *c* (F1pair). The experiment was conducted as follows: First, the Reuter handcrafted solution was taken to be an ideal clustering, and evaluated with all four measures of quality. Then, random errors were incrementally introduced in the ideal clustering. Between each of the 1000 error addition passes, the resulting solution was evaluated with the four F1 methods. We did the same for HD49. Figure 1 shows the deltas between F1pair and F1 best and the text categorization F1 values for Reuter.

Visibly, clustering F1 does not behave like the macro-averaged text categorization F1, as we expected. F1best is slightly optimistic compared to micro-averaged text categorization F1, but always by less than 0.08. F1pair, contrary to our expectations also inflates quality when the amount of error becomes lower than 0.35. F1pair can take values that are inflated or not compared to text categorization F1. Given that unpredictability, F1best will better serve the needs of clustering evaluation. Comparison with text categorization micro-averaged F1 therefore have to account for up to +0.08 inflation for Reuter. For HD49, F1best inflation amounts to a maximum of 0.06 and F1pair is always pessimistic, down to a maximum of 0.24 (not shown).

Although not exactly equivalent due to different ways of counting "misclassifications", F1best and text categorization F1 are nevertheless conceptually related. Considering the benefits of knowing even approximately how well unsupervised text clustering does compared to the upper-bound that supervised text categorization represents outweighs in our opinion considerations of

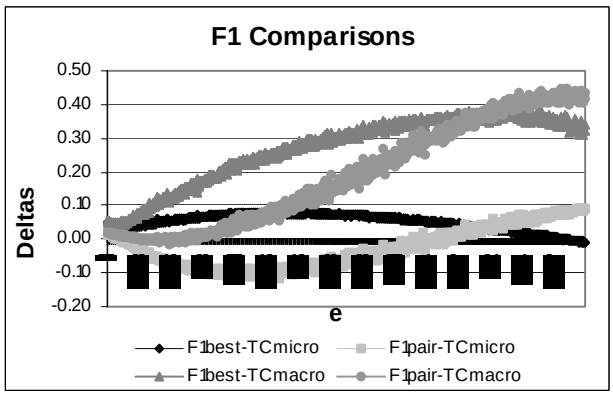

Figure 1. Differences between the two clustering F1 measures (F1pair and F1best) and text categorization F1 in function of random error e (e=0.3 means there is 30% probability error).

mathematical non-equivalence. Indeed, since we have established precisely the correspondence between F1best and text categorization F1 and thus know the maximum level of divergence between these various measures, we can use the F1 measures adapted to text clustering confidently and knowingly, and perform a comparison with text categorization published results with F1 micro-average.

5. Experimental Trial

We now briefly experiment with our methodology to demonstrate its applicability. We evaluate three clustering algorithms using the methodology we proposed in this paper. The first algorithm is the Adaptive Resonance Theory (ART) neural network [8], which is known for its ability to efficiently cluster large, dynamic data sets in an incremental fashion. The second algorithm is the well-known k-means [26] and the third spherical k-means [27] designed to cluster high-dimensional, sparse data points such as text, but normalized to lie on a unit hyper-sphere.

The standard bag-of-words binary vector space representation was used to represent documents [2]. The Reuter training and discarded data sets were used to accumulate word frequency statistics and build the collection vocabulary. Stop words were removed. The only words kept in the test set were those that occur in more than 296 documents in the training and discarded sets. This aggressive feature reduction resulted in a final dimensionality of 598 words-features. The resulting quality is plotted in figure 2. K-means and spherical k-means were used in incremental mode (ART being incremental) with k=93. Results with the two k-means are averaged over 10 trials with random centroids initializations. k cannot be specified for ART, so the vigilance parameter was increased until the number of clusters reached 94 (which was the closest we could get to the expected number of clusters). We observe that spherical k-means is only marginally better than k-means

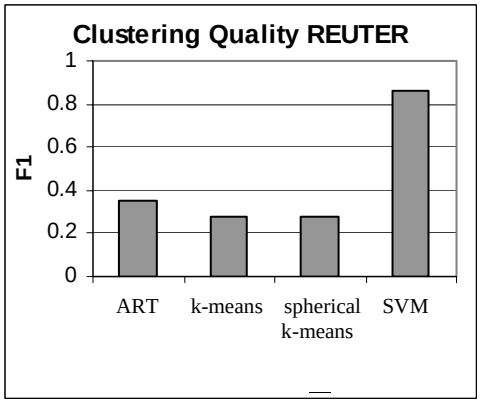

Figure 2. Reuter clustering quality comparison.

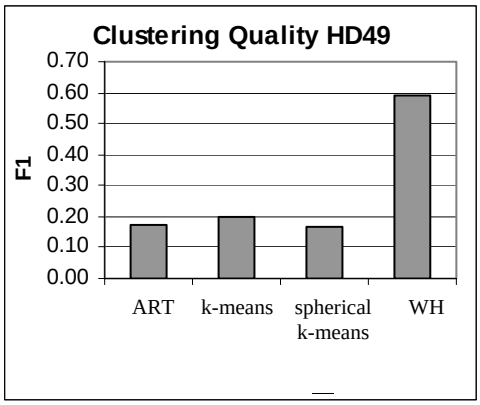

Figure 3. HD49 clustering quality comparison.

and still far to reach the supervised SVM quality. ART is only slightly better, attaining up to 42% of supervised quality. Since F1best values for Reuter clustering can be up to 0.08 inflated, ART may be achieving as low as 28% of supervised quality. Supervised SVM (Support Vector Machine) results are from [7], which is as far as we know the best published results on this data set.

The HD49 documents are pre-processed similarly to Reuter. After aggressive feature reduction, 150 words-features are left. Figure 3 shows the results. In the case of HD49, k-means does slightly better than the other two clustering algorithms. In this case, we compare to WH (Widrow-Hoff) because it is that supervised method for which the best micro-averaged F1 quality on HD49 is reported [17]. K-means reaches 24-34% of the supervised quality (F1best can be up to 0.06 inflated).

We do not claim that these are extraordinary results, but merely that 1) this information becomes extremely valuable for a project manager that needs to decide between supervised and unsupervised approaches to organize vast document archives; and, 2) being able to clearly express clustering quality results in a standard manner with benchmark data will help develop better unsupervised text organisation algorithms that may one day approach supervised quality. Hence, our proposed experimental methodology makes it clear what level of quality is achieved by each algorithm. Then, other investigators could test more advanced feature selection with the same algorithms or other algorithms on the same data sets. The primary advantage is that these other experiments can then compare directly with the results presented here or in other papers using the common methodology. Thus, one can immediately determine what works best and in what circumstances.

6. Conclusion

We had two objectives: first, to improve upon the existing situation whereas most experimental results in text clustering use different data sets and evaluation methodologies. This leads to experimental results that cannot be compared, and hence slows down and even limits scientific development in the field of text clustering. Second, since text clustering can be seen as a low cost alternative to supervised classification, we need means to compare both approaches and decide which one would be best in a particular situation.

To achieve these objectives and solve the related problems, we have proposed a simple, easy to use evaluation framework for text clustering that allows one to compare various text clustering results. Furthermore, the experimental methodology we suggest allows one to compare text clustering with supervised text classification. This provides a useful tool for organizations considering the implementation of a document organization system and hesitating between supervised approaches or the lower cost clustering.

We have demonstrated the use of our experimental framework with three clustering algorithms. Our results show how well these clustering algorithms do compared to supervised text categorization on the exact same data. They reached 24-40% of the quality obtained with classification without any fancy preparation, no labeling, no previously known topics and no training.

There are certainly many other text collections and evaluation methods. Our proposal is based on the widespread use of Reuter and HD49 text collections, as well as of F1 quality measure in supervised text categorization. Experimentations with these data sets and this quality measure have lead to a better understanding of the fundamental issues in the field of text categorization. We hope the same can be achieved for text clustering.

References:

[1] Jain, A.K., Murty, M.N. and Flynn, P.J. (1999). "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, Sept 1999.

[2] Sebastiani, F (2002). "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, Vol. 34, No. 1, March 2002, pp. 1–47.

[3] Attardi, G., Gullì, A. & Sebastiani, F. (1999) Theseus: Categorization by context, In *8th Word Wide Web Conference*, Toronto, Canada, 1999.

[4] Cutting, D. Karger, D., Pedersen, J. and Tukey, J. (1992). "Scatter-gather: A cluster-based approach to browsing large document collections," in *Proceedings of SIGIR'92*, 1992.

[5] Kohonen, T. , Lagus, K. Salojärvi, J, Honkela, J, Paatero, V. and Saarela, A. (2000). "Self Organization of a Document Collection", *IEEE Transactions On Neural Networks*, Vol 11, No. 3, May 2000.

[6] Steinbach, M., Karypis, G., and Kumar, V. (2000). "A Comparison of Document Clustering Techniques", *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 20-23, 2000, Boston, MA, USA.

[7] Yang, Y. & Liu, X. (1999). "A re-examination of text categorization methods". In: *Proceedings of Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, 42-49.

[8] Massey, L. (2003). On the quality of ART1 text clustering. *Neural Networks (16)*5-6 pp. 771-778.

[9] Merkl, D. Text data mining. In *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text.* Marcel Dekker, New York, 1998.

[10] Cohen, W. and Singer, Y. Context-Sensitive Learning Methods for Text Categorization. *ACM Transactions on Information Systems*. 17, 2, 141-173, 1998.

[11] Dumais, S. T., Platt, J., Heckerman, D. and Sahami, M.(1998). Inductive learning algorithms and representations for text categorization. *In Proceedings of ACM-CIKM98*, Nov. 1998, pp. 148-155.

[12] Li, H. and Yamanishi, K. "Text Classification Using ESC-Based Decision Lists," In *Proceedings of International Conference on Information & Knowledge Management(CIKM99)*, pp.122-130, 1999.

[13] Weiss, S. M., Apte, C. , Damerau, F. J. Johnson, D. E. Oles, F. J. Goetz, T. and T. Hampp. Maximizing Text-Mining Performance. IEEE Intelligent Systems, July-August 1999.

[14] Banerjee, A. & Ghosh, J. (2003). "Competitive Learning Mechanisms for Scalable, Incremental and Balanced Clustering of Streaming Texts" In *International Joint Conference on Neural Networks(IJCNN)*, June 2003.

[15] Hersh, W.R., Buckley, C. , Leone, T. J. & Hickam, D.H. (1994) "OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research". In : *Proceedings of SIGIR 1994*: pp.192-201.

[16] Ruiz, M.E. & Srinivasan, P. (1999). "Hierarchical Neural Networks for Text Classification". In: *Proceedings of SIGIR 1999*.

[17] Yang, Y. (1997). "An Evaluation of statistical approach to text categorization". Technical Report CMU-CS-97-127, Computer Science Department, Carnegie Mellon University, 1997.

[18] Yang, Y. (2001). "A study on thresholding strategies for text categorization". In : *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, pp 137-145, 2001.

[19] Grabmeier, J. & Rudolph, A. Techniques of Cluster Algorithms in Data Mining. *Data Mining and Knowledge Discovery* 6(4): 303-360 (2002).

[20] VanRijsbergen, C. J. (1979). Information Retrieval. London: Butterworths, 1979.

[21] Hussin, M.F. & Kamel, M. Document clustering using hierarchical SOMART neural network. In *International Joint Conference on Neural Networks(IJCNN)*, June 2003.

[22] Larsen, B. & Aone, C. (1999). "Fast and effective text mining using linear-time document clustering", In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999, 16 – 22.

[23] Marwick, A.D. (2001). Knowledge Management Technology, *IBM Systems Journal*, Vol. 40, No. 4, 2001.

[24] Dom, B.E (2001). "An Information-Theoretic External Cluster-Validity Measure", *IBM Research Report*, October 2001.

[25] Milligan, G.W., Soon, S.C. & Sokol, L.M. (1983). "The Effect of Cluster Size, Dimensionality, and the number of Clusters on Recovery of True Cluster Structure", *IEEE Trans. On Pattern Analysis and Machine Intelligence* 5(1), Jan. 1983.

[26] MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations", in *Proceedings of the 5th Berkeley symposium on Mathematical Statistics and Probability*. Vol 1, Statistics. Edited by L.M. Le Cam and J. Neyman. Univ of California Press.

[27] Dhillon, I.S. & Modha, D.S. Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*(1):143-175, January 2001.