PERGAMON

AVAILABLE AT
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT

Neural Networks 16 (2003) 771–778

Neural Networks

www.elsevier.com/locate/neunet

2003 Special issue

On the quality of ART1 text clustering

Louis Massey

Royal Military College of Canada, Kingston, Ont., Canada K7K 7B4

Abstract

There is a large and continually growing quantity of electronic text available, which contain essential human and organization knowledge. An important research endeavor is to study and develop better ways to access this knowledge. Text clustering is a popular approach to automatically organize textual document collections by topics to help users find the information they need. Adaptive Resonance Theory (ART) neural networks possess several interesting properties that make them appealing in the area of text clustering. Although ART has been used in several research works as a text clustering tool, the level of quality of the resulting document clusters has not been clearly established yet. In this paper, we present experimental results with binary ART that address this issue by determining how close clustering quality is to an upper bound on clustering quality.

Keywords: Adaptive resonance theory; Text clustering; Text categorization

1. Introduction

We consider the application of clustering to the self-organization of a textual document collection. *Clustering* is the operation by which similar objects are grouped together in an unsupervised manner (Jain, Murty, & Flynn, 1999; Kaufman & Rousseeuw, 1990). Hence, when clustering textual documents, one is hoping to form sets of documents with similar content. Instead of exploring the whole collection of documents, a user can then browse the resulting clusters to identify and retrieve relevant documents. As such, clustering provides a summarized view of the information space by grouping documents by topics. Clustering is often the only viable solution to organize large text collections into topics. The advantage of clustering is realized when a training set and classes 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.

We specifically study text clustering with Adaptive Resonance Theory (ART) (Carpenter & Grossberg, 1995; Grossberg, 1976) neural networks. ART neural networks are known for their ability to perform on-line and incremental clustering of dynamic datasets. Contrary to most other types of artificial neural networks such as the popular Back-propagation Multi-Layer Perceptron (MLP) (Rumelhart, Hinton, & Williams, 1986), ART is unsupervised and allows for plastic yet stable learning. ART detects similarities among data objects, typically data points in an N-dimensional metric space. When novelty is detected, ART adaptively and autonomously creates a new category. Another advantageous and distinguishing feature of ART is its ability to discover patterns at various levels of generality. This is achieved by setting the value of a parameter known as vigilance and denoted by ρ, $\rho \in (0, 1]$. ART stability and plasticity properties as well as its ability to process dynamic data efficiently make it an attractive candidate for clustering large, rapidly changing text collections in real-life environments. Although ART has been investigated previously as a means of clustering text data, due to numerous variations in ART implementations, experimental data sets and quality evaluation methodologies, it is not clear whether ART performs well in this type of application. Since ART seems to be a logical and appealing solution to the rapidly growing amount of textual electronic information processed by organizations, it would be important to eliminate any confusion surrounding the quality of the text clusters it produces. In this paper, we present experimental results with a binary ART neural network (ART1) that address this issue by determining how close clustering quality achieved with ART is to an expected upper bound on clustering quality. We will consider other versions of ART in future work.

E-mail address: massey-l@rmc.ca (L. Massey).

0893-6080/03/$ - see front matter
doi:10.1016/S0893-6080(03)00088-1

2. Related work

We consider one of the many applications of text clustering in the field of Information Retrieval (IR) (VanRijsbergen, 1979), namely clustering that aims at self-organizing textual document collections. This application of text clustering can be seen as a form of classification by topics, hence making it the unsupervised counterpart to Text Categorization (TC) (Sebastiani, 2002). Text self-organization has become increasingly popular due to the availability of large document collections that change rapidly and that are quasi-impossible to organize manually. A typical example is the organization of web documents according to some topics hierarchy like the Yahoo™ hierarchy (Heuser & Rosenstiel, 2000). Even TC becomes unsuitable in such environments because supervised learning of classifiers is not plastic and thus requires retraining upon detection of novelty. Representative work on text clustering includes, among many others, (Cutting, Karger, Pedersen, & Tukey, 1992; Kohonen et al., 2000; Steinbach, Karypis, & Kumar, 2000). There has also been interesting work done on incremental and on-line text clustering with non-neural approaches (see for instance (Can (1993) and Wong and Fu (2000)).

MacLeod and Robertson (1991) were as far as we know the first researchers to consider ART for text clustering. They used a modified version of ART1 in which the similarity computation and weight updates involved domain and task specific knowledge. The inclusion of this knowledge makes the ART implementation more complex but may advantage it over a basic form of ART1. We intend to test this type of ART network in future work, but for now we are interested in establishing the baseline quality that can be achieved with a more basic implementation. The relatively small Keen and Cranfield text collections (800 and 1400 documents, respectively), were used to test MacLeod's algorithm. ART clustering quality was evaluated with the F measure computed on the results of provided queries. This approach to cluster evaluation does not allow for a comprehensive evaluation of the resulting cluster structure, since it only considers the set of queries. The clustering quality results were not very good with $F_1 = 0.25$ and 0.15 (minimum quality value is 0 and maximum 1), but were comparable with other non-neural clustering results published on the same data sets. Merkl (1995) compared ART with Self-Organizing Maps (SOM) (Kohonen, 2001) for the clustering of a small collection of documents and concluded that SOM forms better clusters based on a visual qualitative evaluation. We want to avoid such a subjective evaluation. Moreover, we think that SOM has several weaknesses compared to ART that make it unsuitable for document clustering in a real-life environment characterized by high volume of dynamic data. Indeed, it is unstable under growth while ART provides inherent stability and plasticity. Furthermore, the multiple iterations required to attain convergence with SOM are incompatible with a real-time environment.

On the other hand, some research concluded that ART text clustering resulted in good quality clusters. Vlajic and Card (1998) used a modified ART2 network to create a hierarchical clustering of a small number of web pages. They report that clustering was 'appropriate in all cases, when compared to human performance [...]', but provide no quantitative result. In previous work, we also considered hierarchical clustering, but with ART1 and with a small database of document titles (Massey, 2002a). Clustering quality was deemed adequate based on the percentage of overlap between clusters and expected classification. Our evaluation measure and text collection were non-standard. Rajaraman and Tan (2001) apply fuzzy ART to the task of Topics Detection and Tracking, in which the aspect of topic detection is performed by ART text clustering. They use a small set of 1468 news articles and attempt to measure the ability of the ART network to detect novelty. However, once again only qualitative results are presented. Finally, Kondadadi and Kozma (2002) compare KMART, their soft clustering version of ART, to Fuzzy-ART and k-means. The text data consists of 2000 documents downloaded from the web as well as another 2000 newsgroup documents. Quality evaluation is based on a one-to-one match of the documents in the clusters with the documents in the specified category. K-MART and Fuzzy-ART results are encouraging with above 50% matching for 100–500 document subsets of the original text collections, while *k*-means stays in the range of 22–35% matching. However, this evaluation method is optimistic since it does not account for false positives (i.e. documents that are present in a cluster but do not match documents in the corresponding desired topic).

Our interest in this paper is the study of unsupervised text organization, hence supervised versions of ART (Carpenter Grossberg, & Reynolds, 1991a) applied to text categorization (Petridis, Kaburlasos, Fragkou, & Kehagias, 2001) will not be considered.

3. Experimental settings

We selected two well-established cluster quality evaluation measures: Jaccard (JAC) (Downton & Brennan, 1980) and Fowlkes–Mallows (FM) (Fowlkes & Mallows, 1983):

$$\mathrm{JAC} = a/(a+b+c) \tag{1}$$

$$\mathrm{FM} = a/((a+b)(a+c))^{1/2} \tag{2}$$

where

a is the pair-wise number of true positives, i.e. the total number of document pairs grouped together in the expected solution and that are indeed clustered together by the clustering algorithm;

b is the pair-wise number of false positives, i.e. the number of document pairs not expected to be grouped together but that are clustered together by the clustering algorithm;

c is the pair-wise number of false negatives, i.e. the number of document pairs expected to be grouped together but that are not clustered together by the clustering algorithm.

We also use a measure that computes the F1 clustering quality value. It uses the same underlying pair-wise counting procedure as Jaccard and Fowlkes–Mallow to establish a count of false negatives and false positives, but combines those values following the F-measure (VanRijsbergen, 1979) formulae:

$$F_\beta = (\beta^2 + 1)pr/[\beta^2 p + r] \tag{3}$$

where $p = a/(a+b)$ is known as the precision and $r = a/(a+c)$ as recall. β is set to 1 to give equal weighting to precision and recall. We must note that other text clustering work, such as (Larsen and Aone (1999) and Wong and Fu (2000)), have evaluated clustering quality with F_1 computed on the best cluster-class match. Our initial experiments with this approach indicated that it might unfairly inflate quality. We have not conducted an indepth analysis of this divergence at this point in time, but we will as part of future work.

The 'ModApte' split (Apte, Fred Damerau, & Weiss, 1994) of the Reuter-21578 Distribution 1.0[1] data set is used for our experiments. This data sets is known to be challenging because of skewed class distribution, multiple overlapping categories, and its real-life origin (Reuter newswires during the year 1987, in chronological order). We evaluate clustering results against the desired solution originally specified by Reuter's human classifiers. We only use the desired solution information to evaluate clustering results, i.e. after the clusters have been formed. Reuter is a benchmark data set for TC. Using this data set specific split and the F_1 quality measure makes comparison with published TC results (Yang & Liu, 1999) on the same split possible. This is an important and we believe innovative aspect of our experimental approach: TC F_1 quality results are used as an upper bound for cluster quality since learning in a supervised framework with labeled data provides the best possible automated text classification (with current technology). Thus, clustering can be expected to approach this level of quality but not exceed it since it relies solely on the information present in 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.

We use the k-means (MacQueen, 1967) clustering algorithm to establish a lower bound for quality. Our rationale is that since k-means represents one of the simplest possible approaches to clustering, one would expect that any slightly more advanced algorithm would exceed its clustering quality. The parameter k is set to the number of topics (93) specified by the domain experts who manually organized the Reuter text collection. K-means initial cluster centroids are determined randomly and clustering results are averaged over 10 trials to smooth out extreme values obtained from good and bad random initialization. Our hope is that ART clusters would exceed significantly the quality obtained with k-means and approach the quality of supervised TC.

In this set of experiments, we use the simplest form of ART, binary ART1 in fast learning mode, to establish what should be the baseline level of quality attainable by ART neural networks. ART1 networks consist of two layers of neurons: N input neurons and M output neurons, where N is the input size and M the number of clusters. Neurons are fully connected with both feed-forward and feedback weighted links. The feed-forward links connecting to the output neuron j are represented by the real vector $\mathbf{W}_j$ while the feedback links from that same neuron are represented by the binary vector $\mathbf{T}_j$. The latter stores the prototype representing cluster j. We specifically use Moore's ART1 implementation (Moore, 1988), as follows:

1. Initialize network weights and provide parameter values:
   $0 < \rho \le 1$ (the vigilance parameter) and $L > 1$
   $\mathbf{W}_j = 1/(1+N)$ for all forward connection weights
   $\mathbf{T}_j = 1$ for all feedback connection weights
2. Set the output neurons activation $u_j = 0$ for $j = 1 \ldots M$ and present a document $\mathbf{X}_k$ to the network
3. Compute output activations: $u_j = \mathbf{X}_k \cdot \mathbf{W}_j$ for $j = 1 \ldots M$ and where $\cdot$ is the inner product.
4. Competition between output units: select the most similar category represented by output neuron j^* with maximal activation.
5. Vigilance test: determine if j^* is close enough to $\mathbf{X}_k$:
   $\|\mathbf{X}_k \wedge \mathbf{T}_{j^*}\| / \|\mathbf{X}_k\| \ge \rho$
   where '$\wedge$' is the logical AND operation
   If true, go to step 6 (resonance mode); otherwise, go to step 8 (search mode).
6. Update weights for winning node:
   $\mathbf{T}'_{j^*} = \mathbf{T}_{j^*} \wedge \mathbf{X}_k$
   $\mathbf{W}'_{j^*} = \mathbf{L}(\mathbf{T}_{j^*} \wedge \mathbf{X}_k)/(\mathbf{L} - 1 + \|\mathbf{T}_{j^*} \wedge \mathbf{X}_k\|)$
7. Return to step 2 with a new document.
8. $u_{j^*} = -1$ (remove category j^* from current search) and return to step 4.

The vigilance parameter $\rho \in (0, 1]$ determines the level of abstraction at which ART discovers clusters. Moreover, the minimal number of clusters present in the data can be determined by *minimal vigilance* (Massey, 2002b), computed

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

as $\rho_{min} < 1/N$ where N is the number of features (words) used to represent a document. We chose a value of $\rho_{min} = 0.0005$ as the initial vigilance parameter and we increment it until we find the best clustering quality. We stop increasing vigilance when more than 200 clusters are obtained because such a large number of clusters would simply result in information overload for a user and therefore not achieve the intended objective of text clustering.

A binary vector-space (Salton & Lesk, 1968) representation was created for the Reuter ModApté test set. Only the test set was clustered for compatibility reasons with TC results and also because in unsupervised learning, one must assume that a training set is unavailable. A standard stop word list was used to remove frequent words and a simple feature reduction by term selection based on term frequency was applied to reduce the dimensionality of the original documents feature space. This approach was judged very effective for TC by Yang and Pedersen (1997)).

4. Experimental results

We eliminated words that appear in 10, 20, 40 and 60 or less documents. In the first case, a total of 2282 term features were retained while in the last only 466 were. Our experiments indicated that less radical feature selection not only increased the number of features and consequently processing time, but also resulted in lower quality clusters in some cases (Fig. 1). Best quality is achieved at vigilance value of 0.05, with 106 clusters, a number close to the expected number of topics specified by the domain experts who labeled the data (93). Vigilance levels past 0.1 result in over 250 clusters, which is not desirable for users as explained previously.

The results shown in Fig. 1 were obtained with a single pass in the data. In reality, ART converges to a stable representation after at most $N - 1$ presentations of the data (Georgiopoulos, Heileman, & Huang, 1990). By stable, it is meant that if the same document is presented several times to the network, it should be assigned to the same category, and presenting the same inputs over and over should not change the cluster prototype values. Unstable clusters are problematic since an identical document submitted at different times may end up in different clusters. Furthermore, we show in Fig. 2a that cluster quality increases after ART has stabilized. Only four iterations were required to attain a stable representation, which is much less than the theoretical upper bound of $N - 1$. However, in real world, high-volume operations this could still be a problem as little idle time in the system operation may be available to stabilize topics representation.

We have processed the documents in their natural order, i.e. the chronological order in which they have been created and thus the order in which they would be submitted to a classification system. This simulates the real-world environment where there is no control over the order in which documents are created. As with any on-line clustering algorithm, ART gives different results depending on the order of presentation of the data. This is expected since clustering decisions are taken for each sequentially submitted document, compared to batch clustering that considers all data at once. We submitted the data set in 15 different random orders to ART and averaged clustering quality for each order. Fig. 2b shows that other orders of presentation are much worse than the natural, chronological order of the documents in Reuter. This is encouraging because if quality was higher for other orders of presentation, one would face the problem of finding the best order among the very large number of possibilities or design a way to combine results from different orders. It is possible that other text collections face this problem. Maybe it just happens that Reuter natural order is simply compatible with the desired solution, while in other cases this may not happen. After all, there are many ways to organize large text collections. Some versions of ART use a similarity measure claimed to make it less susceptible to order variations (Sadananda & Sudhakara Rao, 1995). Our initial experiments with Sadananda and Sudhakara Rao's similarity measures are inconclusive at this point in time. However, their proposed similarity measure will in some circumstances not allow the vigilance test (step 5 of the algorithm) to pass even for a newly created category node. Hence, novelty integration becomes impossible unless the vigilance

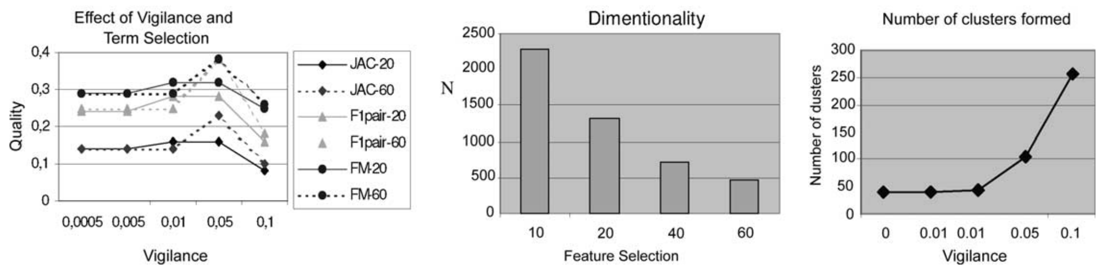

Fig. 1. (a) More radical term selection (removing words appearing in 60 documents or less) results in better clustering quality in some cases, (at vigilance 0.05), compared to removing terms appearing in only 20 documents or less. (b) More radical feature selection also results in much smaller data set dimentionality which in turn allows for more rapid processing. (c) Vigilance 0.05 finds a number of cluster close to the expected number of 93. Vigilance of 0.1 creates too many clusters for users.

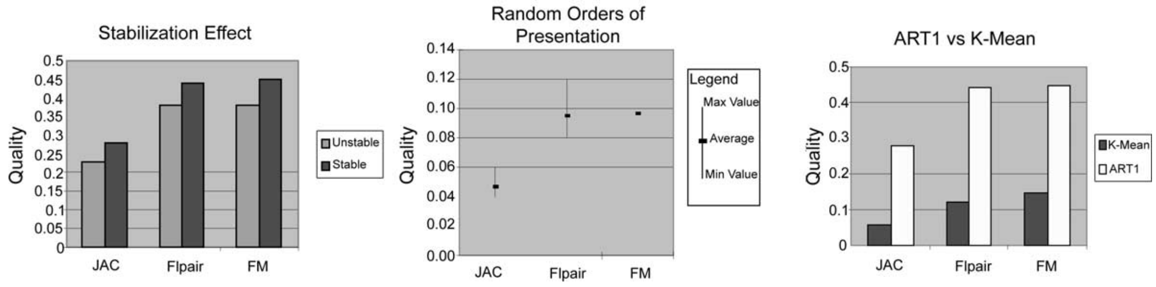

Fig. 2. All results shown for vigilance 0.05. (a) Stabilization improves ART clustering quality. (b) Random orders of presentation even when stabilized give much worse clusters than the natural order of the documents in Reuter. (c) ART clusters (in natural order, stabilized) are of better quality than k-means ($k = 93$).

test is not performed for new category nodes. This amounts to a simple modification to the algorithm listed in this paper.

Fig. 2c. shows that ART1 cluster quality clearly exceeds the lower bound established by K-means. However, clustering quality achieved by ART with random orders of presentation is comparable to K-means. This implies that if the natural order of presentation does not correspond to the chosen organization of the data, ART1 will not do better than K-means. We now compare ART1 clustering quality to the upper bound expected for cluster quality: the best TC results obtained with Support Vector Machines (SVM) and k-Nearest Neighbors (kNN) published in Yang and Liu (1999) (Fig. 3). ART1 achieves 51.2% of the TC quality. Comparing to 16.3% level of quality for k-means, the lower bound, ART1 does much better but is still only about half-way to the optimal expected quality. There is also a potential for lower quality with other orders of presentation. We must however point out that the solution we use to evaluate clusters is merely one among many other useful ways to organize the text collection. Hence, we merely evaluate the ability of ART to recover the specified solution. A users study may better validate the structure discovered by ART, but such studies are costly and also subjective.

We also built a small program that simulates ART cluster prototype updating behavior and attempts to assign each document to its desired topic. We found that 2.2% of documents could not be assigned to their designated topic. Thus, even in the best conditions, perfect clustering is not possible with ART1 with this data set in natural order since right from the start about 2% quality is lost to the order of presentation. This can be explained as follows: a document is assigned to a topic if it has a sufficiently number of overlapping features, as determined by the vigilance parameter. Furthermore, over time as documents are submitted to the network, the number of active features in prototype vectors decreases, which makes matching a document with its desired topic more unlikely. This is caused by the prototype updating mechanism that intersects documents assigned to a topic with that topic prototype. So, in our simulation, if no feature of a document overlap with the prototype for the document desired topic, it means that the expected solution cannot be satisfied with ART1 and with this order of presentation. Better prototype updating may be required to improve quality, such as (MacLeod & Robertson, 1991), which used the union of a document and its prototype features.

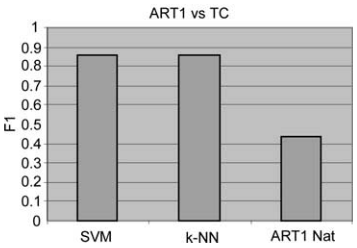

Fig. 3. ART1 clustering F_1 with stabilization at vigilance 0.05 for data in its natural order (ART1 Nat). For both TC methods (SVM and k-NN), the micro-averaged F1 values is used for compatibility with our F1pair measure.

Finally, while conducting our experiments, we noticed an interesting phenomenon: clusters do not necessarily form at the specified level of abstraction. In other words, ART sometimes discovers topics generalizing or specializing desired topics. A generalization is a cluster that includes two or more *classes* (a class being a group of documents representing a desired topic). A specialization on the other hand is a class that includes two or more clusters. In a sense, such behavior can be expected from clustering algorithms since they rely solely on similarity among data items rather than on directions by a domain expert. Fig. 4 shows a portion of the confusion table built from the *matches* between clusters and classes to illustrate generalizations and specializations. A match is the number of documents a class and a cluster have in common. We note from Fig. 4 that:

- cluster 0 is dominated by 169 documents from class 25 (topic earnings) out of a total of 1087 (16%) documents expected to belong to that class;

0	1	2	3	9	10	11	12	14	15	16	17	18	24	25
6	13	3	1		2	1	0	5	0	11	0	0	6	$\boxed{169}$
11	32	7	1		1	9	1	19	1	29	13	9	8	$\boxed{741}$
0	0	1	0		0	0	0	0	0	14	2	2	0	1
97	92	154	26		7	77	5	42	27	$\boxed{598}$	121	88	5	97

Fig. 4. A shortened version of the confusion table: computed clusters 0-3 on rows and desired classes 0–25 on columns. The first row shows the class number. Highlighted values indicate best match.

- cluster 1 is dominated by 741 documents from class 25 (earnings) out of a total possible of 1087 (68%);
- cluster 3 is dominated by 598 documents from class 16 (acquisitions) out of a total possible of 719 (83%), but there is also a strong presence from class 0 (trade), class 1 (grain), class 2 (crude), class 11 (shipping) and class 18 (interest rate).

Hence, one can consider that clusters 0 and 1 actually correspond to class 25 (earnings) with 910 of the 1087 documents, so topic earnings is specialized. Cluster 3 corresponds to classes 16 (acquisitions), class 0 (trade), class 1 (grain), class 2 (crude), class 11 (shipping), and class 18 (interest rate) for a total of 1106 of 1395 documents. So all these classes are generalized by cluster 3. Other class-cluster matches may be deemed to simply lower generalization and specialization quality. We designed a quality evaluation methodology that, contrary to existing clustering evaluation measures, does not penalize generalizations and specializations. For each class, it looks for all clusters that match the class. This allows for the discovery of specializations. Then, for each of the matching cluster, it also looks for all matching class, which accounts for generalizations. Extraneous documents in the latter case are the clustering errors and a F_1 value is computed based on these errors. Fig. 5 shows quality evaluation with this measure, which we call Sub-Graph Dispersion (SGD) because a sub-graph is created for each class when looking for matches. If one considers generalizations and specializations of the expected solution as acceptable, higher quality can be computed at lower vigilance, but the quality of generalizations and specialization decreases as vigilance increases. This method of evaluation needs to be refined before final conclusions on the true impact of learning at different levels of abstraction on clustering quality can be drawn. For instance, all matches are currently considered but some negatively affect quality and should not be included as being part of generalizations or specializations.

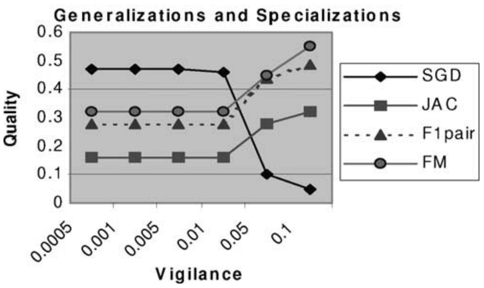

Fig. 5. Increased quality is computed by SGD at lower vigilance by not penalizing generalizations and specializations. Stabilized results shown.

5. Conclusions and future work

Text clustering work conducted with ART up to now has used many different forms of ART-based architectures, as well as different and non-comparable text collections and evaluation methods. This situation resulted in confusion as to the level of clustering quality achievable with ART. As a first step towards resolving this situation, we have tested a simple ART1 network implementation and evaluated its text clustering quality on the benchmark Reuter data set and with the standard F_1 measure. K-means clustering quality was used as the lower bound on quality while published results with supervised TC were used as an upper bound on quality.

Our experiments have demonstrated that text clusters formed by ART1 achieve 51% of TC upper bound and exceed the lower bound considerably. Consequently, about half of the evidence needed to recover the expected document organization solution is available directly from the data under the form of inter-document similarity rather than from costly and time consuming handcrafting of a large labeled training data set. Whether this level of quality is sufficient is a task specific question and ultimately a matter of cost/quality trade-off: it is a choice between higher quality supervised document categorization obtained at high development and maintenance cost versus lower quality clusters obtained at basically no cost. At least we provide here a clear picture of the quality aspect by establishing the baseline quality to be expected with ART. Although ART clusters were of medium quality, ART has the unique advantage of proceeding entirely without human intervention, plus offers the interesting properties of plasticity and stability. Should novelty be detected by the network, a new topic would automatically be created as part of normal system operation. This contrasts with supervised TC, which would require downtime for re-training and related human intervention to prepare a new training set. Therefore, despite lower quality, there may be some situations where ART-based text clustering is a necessity. Furthermore, clustering quality can be increased if one considers discovery of topics at other levels of abstraction as acceptable. Hence, an important area of future research is to explore evaluation measures that do not penalize specializations and generalizations. We are also looking into better feature selection that may also help improve cluster quality. For instance, preliminary experiments with TF-IDF, a well known Information Retrieval measure of term importance, are encouraging in that respect. Moreover, more advanced ART architectures with nonbinary representation (such as ART2 (Carpenter, Grossberg, & Rosen, 1991b), fuzzy ART (Carpenter, Grossberg, & Rosen, 1991c), MacLeod's ART (MacLeod & Robertson, 1991) and FOSART (Baraldi & Alpaydin, 2002)) may further improve cluster quality. We are currently investigating these avenues. As well, in the Reuter collection, topics are not mutually exclusive, while ART1 clustering is. ART based soft clustering such as with KMART (Kondadadi & Kozma, 2002) will thus be explored as yet another possible way to improve clustering.

ART stability and plasticity properties as well as its ability to process dynamic data efficiently make it an attractive candidate for clustering large, rapidly changing text collections in real-life environments. However, in this paper we have only evaluated the static clustering case with some glimpses at the issues that may arise in a more realistic, dynamic environment. For instance, there is a requirement for idle time to allow for stabilization in a text clustering system. We are currently conducting a full assessment of ART's text clustering performance in a simulated realistic environment.

Comparison with other clustering methods has not been our objective. We rather focused on establishing the level of quality achieved by ART within the range defined by a lower and an upper bound. We believe this gives a better appreciation of the level of quality by setting it in a wider framework. Some investigators have evaluated clustering quality with other algorithms on Reuter-21578 and with the F_1 measure, but have used non-standard splits (Larsen & Aone, 1999; Steinbach et al., 2000). So our results cannot be compared directly with theirs. We plan to eventually evaluate other clustering methodologies—particularly incremental clustering algorithms—to compare their clustering quality and ability to function in a dynamic environment to ART's. As well, testing on other text collections is needed to verify if quality results apply to document sets displaying various characteristics.

References

Apte, C., Fred Damerau, F., & Weiss, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, *12*(2), 233–251.

Baraldi, A., & Alpaydin, E. (2002). Constructive feedforward art clustering networks—Part II. *IEEE Transactions on Neural Networks*, *13*, 3.

Can, F. (1993). Incremental clustering for dynamic information, processing. *ACM Transactions on Information Systems*, *11*(2), 143–164.

Carpenter, G. A., & Grossberg, S (1995). Adaptive Resonance Theory (ART). In Handbook of brain theory and neural networks, Arbib, M.A.: MIT Press.

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991a). Artmap: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, *4*, 565–588.

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991b). Art2-a: an adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, *4*, 493–504.

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991c). Fuzzy art: fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, *4*, 759–771.

Cutting, D. Karger, D., Pedersen, J., & Tukey, J (1992). Scatter-gather: a cluster-based approach to browsing large document collections. In *Proceedings of SIGIR'92*.

Downton, M., & Brennan, T (1980). Comparing classifications: an evaluation of several coefficient of partition agreement. In *Proceedings of the meeting of the classification society, Boulder, CO*.

Fowlkes, E., & Mallows, C. (1983). A method for comparing two hierarchical clusterings. *Journal of American Statistical Association*, *78*, 553–569.

Georgiopoulos, M., Heileman, G. L., & Huang, J. (1990). Convergence properties of learning in ART1. *Neural Computation*, *2*(4), 502–509.

Grossberg, S. (1976). Adaptive pattern classification and universal recording: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Heuser, U., & Rosenstiel, W (2000). Automatic construction of local internet directories using hierarchical radius-based competitive learning. In *Proceedings of the fourth world multiconference on systemics, cybernetics and informatics (SCI 2000), Orlando/Florida, Vol. IV (Comunications Systems and Networks)* (pp. 436–441).

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*, 3.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley/Interscience.

Kohonen, T. (2001). *Self-organizing maps*. *Springer series in information*.

Kohonen, T., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a document collection. *IEEE Transactions on Neural Networks*, *11*, 3.

Kondadadi, R., & Kozma, R (2002). A modified fuzzy art for soft document clustering. In *Proceedings of the international joint conference on neural networks, Honolulu, HA*.

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* (pp. 16–22).

MacLeod, K. J., & Robertson, W. (1991). A neural algorithm for document clustering. *Information Processing and Management*, *27*(4), 337–346.

MacQueen, J (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1, Statistics*.

Massey, L (2002a). Structure discovery in text collections. In *Proceedings of the sixth international conference on knowledge-based intelligent information & engineering systems (KES'2002), Italy*.

Massey, L (2002b). Determination of clustering tendency with art neural networks. In *Proceedings of recent adavances in soft-computing (RASC02), Nottingham, UK*.

Merkl, D (1995). Content-based software classification by selforganization. In *Proceedings of the IEEE international conference on neural networks (ICNN'95), Perth, Australia* (pp. 1086–1091).

Moore, B (1988). ART and pattern clustering. In *Proceedings of the 1988 connectionist models summer school* (pp. 174–183).

Petridis, V., Kaburlasos, V. G., Fragkou, P., & Kehagias, A (2001). Text classification using the sigma-FLNMAP neural network. In *Proceedings of international joint conference on neural networks, Washington, DC*.

Rajaraman, K., & Tan, A.-H (2001). Topic detection, tracking and trend analysis using self-organizing neural networks. In *Proceedings of the fifth Pacific-Asia Conference on knowledge discovery and data mining (PAKDD'01), Hong Kong* (pp. 102–107).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, & J. L. McClelland (Eds.), (*Vol. 1*) (pp. 318–364). *Parallel distributed processing: Explorations in the microstructure of cognition, Foundations*, Cambridge, MA: MIT Press.

Sadananda, R., & Sudhakara Rao, G. R. M (1995). ART1: model algorithm characterization and alternative similarity metric for the novelty detector. In *Proceedings IEEE international conference on neural networks, Vol. 5* (pp. 2421–2425).

Salton, G., & Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the ACM*, *15*(1), 8–36.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Steinbach, M., Karypis, G., & Kumar, V (2000). A comparison of document clustering techniques. In *Proceedings sixth ACM SIGKDD* *international conference on knowledge discovery and data mining, Boston, MA*.

VanRijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.

Vlajic, N., & Card, H.-C (1998). Categorizing web pages using modified ART. In *Proceedings of IEEE 1998 Canadian conference on electrical and computer engineering*.

Wong, W.-C., Fu, A., & W.-C (2000). Incremental document clustering for web page classification. In *Proceedings of IEEE 2000 international conference on information society in the 21st century: emerging technologies and new challenges, Japan*.

Yang, Y., & Liu, X (1999). A re-examination of text categorization methods. In *Proceedings of international ACM conference on research and development in information retrieval (SIGIR-99)* (pp. 42–49).

Yang, Y., & Pedersen, J. O (1997). A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th international conference on machine learning, Nashville, USA* (pp. 412–420).