# Can Automatic Personal Categorization deal with User Inconsistency?

Dina Goren-Bar and Tsvi Kuflik

Department of Information Systems Engineering
Ben-Gurion University of the Negev,
P.O.B. 653, Beer-Sheva, 84105 Israel
Phone: 972-8-6472789, Fax: 972-8-6477527
Email:(dinag, tsvikak)@bgumail.bgu.ac.il

**Abstract.** Document categorization is a daily task in every organization, but it is a very subjective process. While automatic document categorization has been widely studied, much challenging research still remains to support user subjective categorization. This study evaluates and compares the application of Self-Organizing Maps (SOM) and Learning Vector Quantization (LVQ) to automatic document classification according to a subjectively predefined set of clusters in a specific domain, and assesses the effect of user inconsistency on this process. Results show that despite the subjective and inconsistent nature of human categorization, automatic document clustering methods correlate well with subjective, personal clustering. Moreover, adapting a system to its users is limited by users' inconsistency, meaning that a perfect adaptation is an impractical goal.

## Introduction

### Motivation

For years now, manual categorization methods have been defined and employed in libraries and other document repositories according to human judgment. Obviously, the categories used by information users are idiosyncratic to the specific user. Hence, a major problem in generating an automatic, adaptable system is to determine to what extent it can reflect the subjective user’s viewpoint, regarding his/her domain of interest. Moreover, users are usually inconsistent about their classifications and have a tendency to change the document classification they use over time [3]. In addition, they may find a document relevant to more than one category, usually choosing just one to host the document. In view of all these problems, the question we address in this paper is how can an automatic clustering system “guess” the user's subjective classification?

The motivation for the current study was to evaluate the possible use of automatic categorization techniques to support manual categorization in a company that performs this task on a daily basis by human experts. In our experiment, the system performed a task similar to one of human categorizers, based on his data. In this study we did not use existing well-defined sets of categorized documents, like Routers or TREC, as they do not reflect the specific subjective user's point of view. Rather, we used a set of manually categorized financial news items, and trained Self-Organizing Maps (SOM) and Learning Vector Quantization (LVQ) Artificial Neural Nets (ANN) to automatically categorize them. Then we measured the difference between the automatically generated set of clusters (or categories) and the pre-defined manual clusters. Our assumption was that if we find only a small difference, we can use a trained ANN to automatically cluster the user's documents.

The study specifically addresses the following questions:

1. To what extent can automatic categorization represent personal, subjective user categorization?
2. What is the effect of the training set size on automatic categorization performance?
3. What is the difference between supervised (LVQ) and unsupervised (SOM) training on the above questions problems?
4. How does user inconsistency affect automatic clustering results?

## Background

Clustering is defined as unsupervised classification of patterns into groups. A wide variety of clustering methods have been presented and applied in a variety of domains, such as image segmentation, object and character recognition, data mining, and information retrieval [6]. One well-known clustering method implements Artificial Neural Nets (ANN).

ANN is a network of simple mathematical "neurons" that are connected by weighted links. The ANN is trained by adjusting the weights between the "neurons." Information retrieval and information filtering are among the various applications where ANN has been successfully tested [2].

There are two main branches of ANN which are distinguished by whether their training method is supervised or unsupervised:

1. The supervised ANN uses a "teacher" to train the model. An error is defined as the difference between the model outputs and the known (expected) outputs. The error back-propagation algorithm adjusts the model connection-weights to decrease the error by repeated presentation of inputs.
2. The unsupervised ANN tries to find clusters of similar inputs when no previous knowledge exists about the number of the desired clusters.

In both cases, once the ANN is trained, it is verified by analyzing inputs not used for the training (a test set).

The SOM, a specific kind of ANN, is a tool that is used for the purpose of automatic document categorization [5],[6],[7]. The SOM is an unsupervised competitive ANN that transforms highly dimensional data to a two-dimensional grid, while preserving the data topology by mapping similar data items to the same cell on the grid (or to neighboring cells). A typical SOM is made up of a vector of nodes for input, an array of nodes as an output map, and a matrix of connections between each output unit and all the input units. Thus, each vector of the input dimension can be mapped to a specific unit on a two-dimensional map. In our case, each vector represents a document, while the output unit represents the category that the document is assigned to.

The LVQ algorithm is a supervised competitive ANN that is closely related to the SOM algorithm. Like the SOM, the LVQ transforms high dimensional data to a two-dimensional grid, but without taking into account data topology. To facilitate the two-dimensional transformation, LVQ uses pre-assigned cluster labels to data items, thus minimizing the average expected misclassification probability. However, unlike the SOM, where clusters are generated automatically based on item similarities, the clusters are predefined. In our case, the cluster labels represent the subjective categorization of the various documents supplied by the user. LVQ training is somewhat similar to SOM training, but it requires that each output unit receive a cluster label a priori to training [7].

To use a clustering mechanism, such as an ANN-based approach, an appropriate document representation is required. One popular model is the vector space model in which a document is represented by a weighted vector of terms [1]. This model suggests that a document may be represented by all meaningful terms included in it. A weight assigned to a term represents the relative importance of that term. One common approach for term weighting is TF ("term frequency") where each term is assigned a weight according to its frequency in the document [9].

## Related work

The need for personalized categorization has been felt for quite some time now, and a great deal of work has been done in various application areas. For example, mail filing assistants, such as the MailCat system [10], proposes folders for email that are similar to the categories or labeling processes performed by the ANN. The MailCat system provides the user with categories (folders) from which he chooses the one he deems most appropriate (every folder has a centroid representative vector).

Clustering methods can define similar groups of documents among huge collections. When clustering is employed on the result of a search engine, it may enhance and ease browsing and selecting relevant information. Zamir & Etzioni [12] evaluated various clustering algorithms and showed precision improvement of the initial search results that varied from 0.1- 0.4 to 0.4 – 0.7. Rauber & Merkl [8] showed that clustering which is applied to general collections of documents could result in an ordered collection of documents grouped into similar groups that is easily browsable by users. Other studies, that used techniques resembling the "one button" (one category) implementation of MailCat, reported similar precision results [10].

The main difference between MailCat and the method we propose is that MailCat is based on specific algorithms developed to achieve "reasonable accuracy" in order to support and adapt specific user interests. In our case, we use a well-known categorization method to represent subjective, personal behavior, which may have interesting implications about the generalization of results to different domains.

The rest of the paper is structured as follows: Next we describe the first experiment performed for evaluating the clustering methods, followed by the experimental results. Afterwards, we present a second experiment, analyze the errors of the first experiment, and describe the effect of user inconsistency on subjective classification. We conclude with a discussion of the results and suggestions for future work.

## First Experiment

### Method

The first experiment was performed in a company that deals with information extraction and categorization. Our purpose was to evaluate possible automation of data items categorization within the company's domain using actual, available data. We chose this environment in order to deal with real-life data in a real-life organization. Therefore, the data collection method resembled a field study. We did not ask the information expert to do an experimental task. Instead, we took a collection comprised of previously clustered items, which was built incrementally as a result of common working procedures. A data set containing 1115 economics and financial data items was used. The data was extracted from an online Internet based source (Yahoo.com). An information specialist read each item and manually clustered the items into one of 15 possible clusters. The size of the cluster varied from 1 to 125 documents per cluster. This approach represents a normal daily operation for information seeking and categorization.

The first experiment compared manual classification with two ANN automatic classification algorithms (SOM and LVQ).

In order to evaluate the ANN performance we used precision and recall. Recall, in a specific category, is calculated as the number of documents that were automatically clustered correctly by the system, divided by the overall number of documents that belong to that cluster originally (given by the expert). Precision is the number of documents automatically clustered correctly, divided by the overall documents clustered by the system to the same category (correct + incorrect).

All documents underwent classical text analysis. "Stop words" were removed. The resulting terms underwent further stemming processing using the Porter stemming algorithm [4]. Finally, normalized weighted term-vectors were generated to represent the documents. In order to reduce the vector length, a low frequency threshold was implemented as a means of dimensionality reduction. Non-informative terms, which appeared in less than 100 documents (less than 10%), were removed. Then, the SOM and LVQ simulations were implemented using the SOM Toolbox for Matlab 5 [11].

Since experimental results may be influenced by the selection of the test and training sets, ten test sets were randomly selected. The size of each was 20% of the overall set (223 items). This size was selected for the test set and the initial training set in order to allow (when available) a sufficient number of examples for each original cluster to be used (an average of 15 documents from each cluster). For each test set, several training sets of different sizes were generated from the remaining 892 items by random selection so that the test set items were never part of the training set.

Training sets sizes comprised 20%, 40%, 60% and 80% of the original data set (223, 446, 669, and 892 items).

The labeling process of the auto-generated clusters consisted of the following process: First, for each automatic cluster (output unit), we verified the original manual classification of each of the documents clustered. Then, we checked the frequency of each manual category within each cluster. Last, the manual cluster with the highest frequency rendered its name to the automatic cluster (output unit).

Two runs were performed, one for SOM and one for LVQ (hence, slight differences were found in the randomly selected test and training sets). SOM and LVQ assigned 0 to 44 items per category in each of the 10 data sets.

**Table 1.** Original data clusters

| set<br>cat | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | 1 | 0\1 | | | | | |
| 2 | | 1\0 | 1\0 | | | | | | | |
| 3 | 0\1 | 0\1 | | | | 0\1 | | 1 | | 1\0 |
| 4 | 0\1 | 0\1 | 0\1 | 1 | 1 | 1 | 1 | 1 | 0\1 | 1 |
| 5 | 2\1 | 2 | 1\3 | 2 | 2\3 | 2 | 2\1 | 2\1 | 2\3 | 1 |
| 6 | 5 | 4\5 | 5\4 | 5 | 5\4 | 5\4 | 4\5 | 5 | 5\4 | 5 |
| 7 | 6 | 6 | 6 | 6 | 7\6 | 6 | 6 | 6 | 6 | 6 |
| 8 | 11 | 11 | 11 | 11 | 10\11 | 11 | 11 | 11 | 11 | 11 |
| 9 | 19\18 | 19\17 | 18\17 | 18 | 19\18 | 20\19 | 18\17 | 18 | 18 | 17\18 |
| 10 | 21\20 | 20 | 22\21 | 18\20 | 19\20 | 18\19 | 20\21 | 19\20 | 21\20 | 21 |
| 11 | 21 | 21 | 21 | 22\21 | 22\21 | 22\21 | 21 | 22 | 22\21 | 21 |
| 12 | 26\27 | 27\26 | 25 | 26 | 27\26 | 27 | 27\26 | 27 | 26 | 26 |
| 13 | 29 | 29 | 30 | 31\29 | 29\30 | 29\30 | 29 | 29 | 29 | 30\29 |
| 14 | 39 | 39\40 | 39\40 | 39 | 38\39 | 38\39 | 40\41 | 38 | 39\40 | 40 |
| 15 | 44 | 44 | 44 | 43\44 | 44\43 | 44 | 44 | 44 | 44 | 43\44 |

*Note:* Categories marked with a "\" indicate the LVQ test set and the SOM test set had different numbers of documents. The number in the left hand side represents the LVQ test set and the number in the right hand side the SOM test set.

## Experiment Results

As shown in Table 1, at some of the test-sets, several original categories contained a minimal number of documents, or no documents at all. Therefore, we calculated an average precision and recall for those categories with more than 10 documents per category (categories 8-15) - see Table 2. It can be seen that recall and precision improve significantly for both LVQ (0.75) and SOM (0.73) for each training set size. We can see a consistent increment in both precision and recall as the training set grows (from 20% to 80%).

**Table 2.** Average LVQ and SOM Precision and Recall results for categories with more than 10 documents.

| Learning set | Percentage | 80% | 60% | 40% | 20% |
|---|---|---|---|---|---|
| | Actual size | 892 | 669 | 446 | 223 |
| **Measure** | **Method** | | | | |
| Recall | LVQ | 0.75 | 0.74 | 0.73 | 0.69 |
| | SOM | 0.73 | 0.71 | 0.68 | 0.62 |
| Precision | LVQ | 0.75 | 0.73 | 0.71 | 0.67 |
| | SOM | 0.73 | 0.70 | 0.70 | 0.65 |

## Second Experiment - User Re-evaluation

### Method

In this experiment we examine the impact of user changing categorization on the automatic clustering. Explicit relevance feedback was requested from the user, with respect to those documents that were misclassified by the system. Since asking the user for feedback on all the data was unfeasible, we concentrated on the items wrongly classified by the system in the first experiment. The information specialist was informed that a portion of the data set was arbitrarily chosen for reclassification. We randomly selected one out of the ten datasets used with LVQ system and one for the SOM system, for which the training set size was 892 messages (80%), for re-classification. In those data sets, 62 messages were wrongly categorized by the LVQ method, and 68 messages were wrongly categorized by the SOM method.

The second experiment is a replica of the first experiment. However, the data set contained only information items that were misclassified by the system in the first experiment.

The user was asked to reclassify the messages without knowing either the original categorization, or the system categorization. The same categories were used as in the first experiment.

### Experiment Results

Table 3 compares the average recall and precision results for the original and the reclassified items for categories containing more than 10 documents per category. Again, there is a noticeable improvement, mainly for LVQ, in comparison to the results from the first experiment. For the LVQ, the average recall increased from 0.72 to 0.81 and average precision increased from 0.76 to 0.86.

**Table 3.** Average LVQ and SOM original and reclassified precision and recall for categories with more than 10 documents.

| **Learning set** | Percentage | Original | Reclassified |
|---|---|---|---|
| **Measure** | **Method** | | |
| Recall | LVQ | 0.72 | 0.81 |
| | SOM | 0.69 | 0.75 |
| Precision | LVQ | 0.76 | 0.86 |
| | SOM | 0.66 | 0.76 |

## Discussion and Future Work

The main purpose of this work was to test the possibility of automating the classification of subjectively categorized data sets. For this we worked with real data gathered from a daily work of information search and categorization.

The overall results confirm the hypothesis that it is possible to automate (with reasonable error), a subjective categorization. Both LVQ and SOM succeeded to learn user's categorization. Performing automatic clustering using either SOM or LVQ provided an overall of 69% to 75% recall and 67% to 75% precision for the two methods. The automatic categorization performance was achieved with a learning set ranging from 223 to 892 documents. This indicates that it is possible to train a system to provide useful results even with a minimal training set.

Another finding is that the supervised learning (LVQ) yields better results than the unsupervised (SOM) method, mainly at the initial steps (69% vs. 62% for recall and 67% vs. 65% for precision for the 223 documents training set). The most surprising part is that the overall performance of supervised and unsupervised ANN is quite similar, given a sufficient number of documents. Our results show that from the initial 10 test sets (Table 1), several categories got eliminated because there wasn't enough data. For most data sets, in categories 1-7 the small number of documents was insufficient for training and therefore yielded incorrect categorization. However, these problematic data sets represent real life.

The results show that recall and precision improves significantly as the size of the training set grows. Table 5 indicates that for LVQ increased from 0.67 to 0.75 and from 0.69 to 0.75 respectively. For the SOM precision and recall increased from 0.62 to 0.73 and from 0.65 to 0.73 respectively.

A major question for the adoption of automatic subjective classification systems is the required size of the training set. If a large training set is needed, it is most likely that users will not use such a system. In our case, the size of the training set varied from ~200 to ~900 data items, an average of 15 to 60 data items per category. While 15 data items may be a reasonable size for a data set, bigger sizes may be problematic. However even initial results for a small training set approximated the overall performance achieved by the larger set.

We conclude from our results that despite the subjective nature of human categorization, an automatic process can resemble subjective categorization, with considerable success.

The reclassification of documents reveals an interesting pattern: About 40% of the documents were classified to their original categories, about one-third was classified according to the system suggestion and the rest got a different, new classification. These findings indicated that given a sufficient training set, the SOM and the LVQ could consistently support user subjective categorization. Moreover, they indicate that trying to get a high degree of correlation between an adaptable system to its users is problematic and might be unnecessary. People change their mind. It is part of human nature. Users might reconsider their classification because they think it was wrong, because some contextual factors influenced their judgment differently than before or because the information item relates to more than one category. Further research will focus on the factors that influence user inconsistency and its impact on "personal" categorization.

## References

1. Baeza-Yates and Ribiero-Neto (1999) *Modern Information Retrieval*, Addison-Wesley, 1999.
2. Boger, Z. Kuflik, T., Shapira, B. and Shoval, P. (2000) Information Filtering and Automatic Keywords Identification by Artificial Neural Networks *Proccedings of the 8th Europian Conference on Information Systems.* pp. 46-52, Vienna, July 2000.
3. Dawes, R. M. (1979). The Robust Beauty of Improper linear models in Decision Making. *American Psychologist*, 34, 571-582.
4. Frakes, W. B., and Baeza-Yates, R. (1992). *Information Retrieval data Structures and Algorithms*, Prentice-Hall.
5. Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1997). WEBSOM - Self-Organizing Maps of Document Collections. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo,* Finland, June 4-6, pages 310-315. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland.
6. Jain, A. K., Murty, M. N., Flynn, P. J. (1999) Data Clustering: A Review, *ACM Computing Surveys*, Vol 31, No. 3 pp. 264-323, September 1999
7. Kohonen, T. (1997). *Self-Organizing Maps*. 2nd ed., Springer-Verlag, Berlin.
8. Rauber A. and Merkl. D. (1999). Using self-organizing maps to organize document archives and to characterize subject matters: How to make a map tell the news of the world *Proceedings of the 10th Intl. Conf. on Database and Expert Systems Applications* (DEXA'99), Florence, Italy.
9. Salton, G., McGill, M. *Introduction to Modern Information Retrieval.* McGraw-Hill New-York (1983).
10. Segal, B. R., and Kephart, J. O., (1999) MailCat: An Intelligent Assistant for Organizing E-Mail, *Proceedings of the Third International Conference on Autonomous Agents*. Pp. 276-282
11. Vesanto, J., Alhoniemi, E., Himberg, J., Kiviluoto, K., & Parviainen, J. (1999). Self-Organizing Map for Data Mining in Matlab: The SOM Toolbox. *Simulation News Europe*, (25):54.
12. Zamir, O., and Etzioni O. (1998), Web Document Clustering: A Feasibility Demonstration, *Proceedings of SIGIR 98*, Melbourne, Australia.