Experiments in Term Weighting and Keyword Extraction in Document Clustering

Christian Borgelt and Andreas Nürnberger
Dept. of Knowledge Processing and Language Engineering
Otto-von-Guericke-University of Magdeburg
Universitätsplatz 2, D-39106 Magdeburg, Germany
{borgelt,nuernb}@iws.cs.uni-magdeburg.de

Abstract

We study methods to initialize or bias different clustering methods using prior information about the "importance" of a keyword w.r.t. the whole document collection or a specific cluster. These studies give us hints on how to initialize clustering methods in order to improve performance if prior knowledge is available. This can be especially useful if a user-specific clustering of a document collection or web search result set is desired. Furthermore, we discuss whether one should draw on information measures to extract keywords that can be used as describing features for document clusters.

1 Introduction

The problem of finding descriptive weights for terms in document collections in order to improve retrieval performance has been studied extensively in the past (see, for instance, [12; 24; 23]). To achieve an improved classification or clustering performance for a given text collection it is usually necessary to select a subset of all describing features (i.e. keywords) and/or to re-weight the features w.r.t. a specific classification or clustering goal. Consequently, several studies were conducted in this direction. For example, it was explored how to select keywords based on statistical and information theoretical measures [9; 21; 28] or how to combine clustering and keyword weighting techniques [10] in order to improve the clustering performance. Nevertheless, it is still unclear to what extend term re-weighting influences the clustering performance and whether initial—global or cluster specific—term re-weighting can be used to bias or improve the clustering performance. Therefore, in the following, we compare clustering with and without term re-weighting techniques using different hard and fuzzy clustering methods. Furthermore, by interpreting the resulting cluster prototypes (cluster centers), we discuss briefly whether one should draw on information theoretical measures to extract keywords that can be used to appropriately describe a cluster.

This paper is organized as follows: In Section 2 we briefly review some basics of fuzzy clustering and transfer fuzzy clustering ideas to learning vector quantization. In Section 3 we review pre-processing methods for documents and in particular the vector space model, which we use to represent documents. In Section 4 we present our experimental results of clustering web page collections using different weighting approaches and finally, in Section 5, we draw conclusions from our discussion.

2 Clustering

It is not difficult to see that classical c-*means clustering* [7; 4] and standard *learning vector quantization* applied to clustering [17; 18] are very similar: a point that one method converges to is a stable point of the other, in particular, if learning vector quantization is applied in batch mode. Since classical c-means clustering has been generalized to fuzzy clustering [1; 2; 14], the idea suggests itself to transfer some ideas that have been developed in fuzzy clustering to competitive learning, with the aim of achieving a higher flexibility. The basic idea of this transfer is that the update of a reference vector in competitive learning can be seen as an exponential decay of information gained from data points processed in earlier steps—a scheme that may just as well be applied to a covariance matrix describing the size and shape of a cluster.

Such online clustering has at least two advantages for the application domain we are concerned with here, that is, for clustering collections of documents. The first is that, due to the fact that the cluster parameters are updated more often, while the greater part of the overhead comes from the computations of the distances between the data points and the cluster centers, it can be faster than standard fuzzy clustering. Secondly, this approach to clustering makes it easier to handle documents that become available in a true online fashion, because updates need only few documents, not the whole collection.

In the following, we briefly review fuzzy clustering and learning vector quantization as we use it in our experiments. For a more detailed discussion and evaluation of these methods for document clustering see [5].

2.1 Fuzzy Clustering

While most classical clustering algorithms assign each datum to exactly one cluster, thus forming a crisp partition of the given data, fuzzy clustering allows for *degrees of membership*, to which a datum belongs to different clusters [1; 2; 14]. Most fuzzy clustering algorithms are objective function based: they determine an optimal (fuzzy) partition of a given data set $\mathbf{X} = \{\vec{x}_j \mid j = 1, \ldots, n\}$ into c clusters by minimizing an objective function

$$J(\mathbf{X}, \mathbf{U}, \mathbf{C}) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{w} d_{ij}^{2}$$

subject to the constraints

$$\sum_{j=1}^{n} u_{ij} > 0, \qquad \text{for all } i \in \{1, \ldots, c\}, \qquad \text{and} \quad (1)$$

$$\sum_{i=1}^{c} u_{ij} = 1, \qquad \text{for all } j \in \{1, \ldots, n\}, \tag{2}$$

where $u_{ij} \in [0, 1]$ is the membership degree of datum $\vec{x}_j$ to cluster i and d_{ij} is the distance between datum $\vec{x}_j$ and cluster i. The $c \times n$ matrix $\mathbf{U} = (u_{ij})$ is called the *fuzzy partition matrix* and $\mathbf{C}$ describes the set of clusters by stating location parameters (i.e. the cluster center) and maybe size and shape parameters for each cluster. The parameter w, $w > 1$, is called the *fuzzifier* or *weighting exponent*. It determines the "fuzziness" of the classification: with higher values for w the boundaries between the clusters become softer, with lower values they get harder. Usually $w = 2$ is chosen. Hard clustering results in the limit for $w \to 1$. However, a hard assignment may also be determined from a fuzzy result by assigning each data point to the cluster to which it has the highest degree of membership.

Constraint (1) guarantees that no cluster is empty and constraint (2) ensures that each datum has the same total influence by requiring that the membership degrees of a datum must add up to 1. Because of the second constraint this approach is usually called *probabilistic fuzzy clustering*, since with it the membership degrees for a datum formally resemble the probabilities of its being a member of the corresponding clusters. The partitioning property of a probabilistic clustering algorithm, which "distributes" the weight of a datum to the different clusters, is due to this constraint.

Unfortunately, the objective function J cannot be minimized directly. Therefore an iterative algorithm is used, which alternately optimizes the membership degrees and the cluster parameters [1; 2; 14]. That is, first the membership degrees are optimized for fixed cluster parameters, then the cluster parameters are optimized for fixed membership degrees. The main advantage of this scheme is that in each of the two steps the optimum can be computed directly. By iterating the two steps the joint optimum is approached (although, of course, it cannot be guaranteed that the global optimum will be reached—the algorithm may get stuck in a local minimum of the objective function J).

The update formulae are derived by simply setting the derivative of the objective function J w.r.t. the parameters to optimize equal to zero (necessary condition for a minimum). Independent of the chosen distance measure we thus obtain the following update formula for the membership degrees [14]:

$$u_{ij} = \frac{d_{ij}^{-\frac{2}{w-1}}}{\sum_{k=1}^{c} d_{kj}^{-\frac{2}{w-1}}}, \tag{3}$$

that is, the membership degrees represent the relative inverse squared distances of a data point to the different cluster centers, which is a very intuitive result.

The update formulae for the cluster parameters, however, depend on what parameters are used to describe a cluster (location, shape, size) and on the chosen distance measure. Therefore a general update formula cannot be given. Here we briefly review the three most common cases: The best-known fuzzy clustering algorithm is the fuzzy c-means algorithm, which is a straightforward generalization of the classical crisp c-means algorithm. It uses only cluster centers for the cluster prototypes and relies on the *Euclidean distance*, i.e.,

$$d_{ij}^2 = d^2(\vec{x}_j, \vec{\mu}_i) = (\vec{x}_j - \vec{\mu}_i)^\top (\vec{x}_j - \vec{\mu}_i),$$

where $\vec{\mu}_i$ is the center of the i-th cluster. Consequently it is restricted to finding spherical clusters of equal size. The resulting update rule is

$$\vec{\mu}_i = \frac{\sum_{j=1}^{n} u_{ij}^w \vec{x}_j}{\sum_{j=1}^{n} u_{ij}^w}, \tag{4}$$

that is, the new cluster center is the weighted mean of the data points assigned to it, which is again a very intuitive result.

The Gustafson-Kessel algorithm [13] uses the *Mahalanobis distance*, i.e.,

$$d_{ij}^2 = d^2(\vec{x}_j, \vec{\mu}_i) = (\vec{x}_j - \vec{\mu}_i)^\top \mathbf{\Sigma}_i^{-1} (\vec{x}_j - \vec{\mu}_i),$$

where $\vec{\mu}_i$ is the cluster center and $\mathbf{\Sigma}_i$ is a cluster-specific covariance matrix with determinant 1 that describes the shape of the cluster, thus allowing for ellipsoidal clusters of equal size. This distance function leads to same update rule (4) for the clusters centers. The covariance matrices are updated according to

$$\mathbf{\Sigma}_i = |\mathbf{\Sigma}_i^*|^{-\frac{1}{m}} \mathbf{\Sigma}_i^* \qquad \text{where} \tag{5}$$
$$\mathbf{\Sigma}_i^* = \frac{\sum_{j=1}^{n} u_{ij}^w (\vec{x}_j - \vec{\mu}_i)(\vec{x}_j - \vec{\mu}_i)^\top}{\sum_{j=1}^{n} u_{ij}^w}$$

and m is the number of dimensions of the data space. $\mathbf{\Sigma}_i^*$ is called the *fuzzy covariance matrix*, which is simply normalized to determinant 1 to meet the abovementioned constraint. Compared to standard statistical estimation procedures, this is also a very intuitive result. It should be noted that the restriction to cluster of equal size may be relaxed by simply allowing general covariance matrices. However, depending on the characteristics of the data, this additional degree of freedom can deteriorate the robustness of the algorithm.

Finally, the fuzzy maximum likelihood estimation (FMLE) algorithm [11] is based on the assumption that the data was sampled from a mixture of c multivariate normal distributions as in the statistical approach of mixture models [8; 3]. It uses a (squared) distance that is inversely proportional to the probability that a datum was generated by the normal distribution associated with a cluster, i.e.,

$$d_{ij}^2 = \left(\frac{\theta_i}{\sqrt{(2\pi)^m |\mathbf{\Sigma}_i|}} \exp\left(-\frac{1}{2} (\vec{x}_j - \vec{\mu}_i)^\top \mathbf{\Sigma}_i^{-1} (\vec{x}_j - \vec{\mu}_i) \right) \right)^{-1}$$

where θ_i is the prior probability of the cluster, $\vec{\mu}_i$ is the cluster center, $\mathbf{\Sigma}_i$ a cluster-specific covariance matrix, which in this case is not required to be normalized to determinant 1, and m the number of dimensions of the data space. For the FMLE algorithm the update rules are not derived from the objective function due to technical obstacles, but by comparing it to the well-known expectation maximization (EM) algorithm [6] for a mixture of normal distributions [8; 3], which, by analogy, leads to the same update rules for the cluster center and the cluster-specific covariance matrix [14]. The prior probability is, in direct analogy to statistical estimation, computed as

$$\theta_i = \frac{1}{n} \sum_{j=1}^{n} u_{ij}^w. \tag{6}$$

Since the high number of free parameters of the FMLE algorithm renders it unstable on certain data sets, it is usually recommended [14] to initialize it with a few steps of

the very robust fuzzy c-means algorithm. The same holds, though to a somewhat lesser degree, for the Gustafson-Kessel algorithm.

It is worth noting that of both the Gustafson-Kessel as well as the FMLE algorithm there exist so-called *axes-parallel* versions, which restrict the covariance matrices $\mathbf{\Sigma}_i$ to diagonal matrices and thus allow only axes-parallel ellipsoids [15]. These variants have certain advantages w.r.t. robustness and execution time.

2.2 Learning Vector Quantization

Learning vector quantization [17; 18], in its classical form, is a competitive learning algorithm that has been developed in the area of artificial neural networks and that can be applied to classified as well as unclassified data. Here we confine ourselves to unclassified data, where the algorithm consists in iteratively updating a set of c so-called *reference vectors* $\vec{\mu}_i$, $i = 1, \ldots, c$, each of which is represented by a neuron. For each data point $\vec{x}_j$, $j = 1, \ldots, n$, the closest reference vector (the so-called "winner neuron") is determined and then this reference vector (and only this vector) is updated according to

$$\vec{\mu}_i^{(\mathrm{new})} = \vec{\mu}_i^{(\mathrm{old})} + \eta_1 \left(\vec{x}_j - \vec{\mu}_i^{(\mathrm{old})} \right), \tag{7}$$

where η_1 is a learning rate. This learning rate usually decreases with time in order to avoid oscillations and to enforce the convergence of the algorithm.

Membership degrees can be introduced into this basic algorithm in two different ways. In the first place, one may employ an activation function for the neurons, for which a radial function like the

$$\text{Cauchy function} \qquad f(r) = \frac{1}{1 + r^2} \qquad \text{or the}$$

$$\text{Gaussian function} \qquad f(r) = e^{-\frac{1}{2} r^2}$$

may be chosen, where r is the (radial) distance from the reference vector. In this case all reference vectors are updated for each data point, with the update being weighted with the value of the activation function. However, this scheme, which is closely related to *possibilistic fuzzy clustering* [19], usually leads to unsatisfactory results, since there is no dependence between the clusters, so that they tend to end up at the center of gravity of all data points. This corresponds to the fact that in possibilistic fuzzy clustering the objective function is truly minimized only if all cluster centers are identical [27]. Useful results are obtained only if the method gets stuck in a local minimum, which is an undesirable situation.

An alternative is to rely on a normalization scheme as in *probabilistic fuzzy clustering*, that is, to compute the weight for the update of a reference vector as the relative inverse (squared) distance from this vector (cf. the computation of the membership degrees in fuzzy clustering, see formula (3)), or as the *relative* activation of a neuron. This is the approach we employ here, that is, we use

$$\vec{\mu}_i^{(\mathrm{new})} = \vec{\mu}_i^{(\mathrm{old})} + \eta_1 u_{ij} \left(\vec{x}_j - \vec{\mu}_i^{(\mathrm{old})} \right) \tag{8}$$

with u_{ij} defined as in equation (3). Furthermore we associate with each neuron not only a reference vector $\vec{\mu}_i$, but also a covariance matrix $\mathbf{\Sigma}_i$, which describes the shape and (if we do not require it to be normalized to determinant 1) the size of the represented cluster.

In order to find an update rule for this covariance matrix, we observe that the above equation (7) may also be written as

$$\vec{\mu}_i^{(\mathrm{new})} = (1 - \eta_1)\, \vec{\mu}_i^{(\mathrm{old})} + \eta_1\, \vec{x}_j,$$

which shows that the update can be seen as an exponential decay of information gained from data points processed earlier. Transferring this idea to the covariance matrices $\mathbf{\Sigma}_i$ and drawing on equation (5) leads directly to

$$\mathbf{\Sigma}_i^{(\mathrm{new})} = (1 - \eta_2)\, \mathbf{\Sigma}_i^{(\mathrm{old})} + \eta_2\, (\vec{x}_j - \vec{\mu}_i)(\vec{x}_j - \vec{\mu}_i)^\top, \tag{9}$$

where η_2 is a learning rate, which, in general, differs from the learning rate η_1 for the reference vectors. In the fuzzy case this update may be weighted, as the update of the reference vectors, by the relative inverse (squared) distance of the data point from the reference vector or by the relative neuron activation.

It should be noted that versions of this algorithm that require the covariance matrix to be normalized to determinant 1 or restrict the covariance matrix to a diagonal matrix may be considered, too. Such constraints can improve the robustness or the execution time of the algorithm. Furthermore it should be noted that the updates may be executed in batch mode, aggregating the changes resulting from the data points and actually updating the reference vectors and covariance matrices only at the end of an epoch.

Finally, it should be noted that the additional elements of the FMLE algorithm may also be transferred to a competitive learning scheme, by using the reciprocal of the special distance function as the activation function of the neurons. The additional parameter θ_i, that is, the weight or prior probability of a cluster, is then updated according to

$$\theta_i^{(\mathrm{new})} = (1 - \eta_3)\, \theta_i^{(\mathrm{old})} + \eta_3\, u_{ij},$$

where η_3 is another learning rate (different from η_1 and η_2) and u_{ij} is defined as in equation (3). However, in the application described below, we confine ourselves to the simpler approach discussed above, which in addition to membership degrees introduces only covariance matrices.

3 Clustering Document Collections

To be able to cluster text document collections with the methods discussed above, we have to map the text files to numerical feature vectors. Therefore, we first applied standard preprocessing methods, i.e., stopword filtering and stemming (using the Porter Stemmer [22]), encoded each document using the vector space model [23] and finally selected a subset of terms as features for the clustering process as briefly described in the following.

3.1 The Vector Space Model

The vector space model represents text documents as vectors in an m-dimensional space, i.e., each document j is described by a numerical feature vector $\vec{x}_j = (x_{j1}, \ldots, x_{jm})$. Each element of the vector represents a word of the document collection, i.e., the size of the vector is defined by the number of words of the complete document collection.

For a given document j the so-called weight x_{jk} defines the importance of the word k in this document with respect to the given document collection C. Large weights are assigned to terms that are frequent in relevant documents but rare in the whole document collection [24]. Thus a weight x_{jk} for a term k in document j is computed as the term frequency tf_{jk} times the inverse document frequency idf_k,

which describes the term specificity within the document collection.

In [25] a weighting scheme was proposed that has meanwhile proven its usability in practice. Besides term frequency and inverse document frequency (defined as $\mathrm{idf}_k = \log(n/n_k)$), a length normalization factor is used to ensure that all documents have equal chances of being retrieved independent of their lengths:

$$x_{jk} = \frac{\mathrm{tf}_{jk} \log \frac{n}{n_k}}{\sqrt{\sum_{l=1}^{m} \left(\mathrm{tf}_{jl} \log \frac{n}{n_l}\right)^2}}, \tag{10}$$

where n is the size of the document collection C, n_k the number of documents in C that contain term k, and m the number of terms that are considered.

Based on a weighting scheme a document j is described by an m-dimensional vector $\vec{x}_j = (x_{j1}, \ldots, x_{jm})$ of term weights and the similarity S of two documents (or the similarity of a document and a query vector) can be computed based on the inner product of the vectors (by which—if we assume normalized vectors—the cosine between the two document vectors is computed), i.e.

$$S(\vec{x}_j, \vec{x}_k) = \sum_{l=1}^{m} x_{jl} \cdot x_{kl}. \tag{11}$$

For a more detailed discussion of the vector space model and weighting schemes see, for instance, [12; 24; 23].

Note that for normalized vectors the scalar product is not much different in behavior from the Euclidean distance, since for two vectors $\vec{x}$ and $\vec{y}$ it is

$$\cos\varphi = \frac{\vec{x}\vec{y}}{|\vec{x}| \cdot |\vec{y}|} = 1 - \frac{1}{2}\, d^2\left(\frac{\vec{x}}{|\vec{x}|}, \frac{\vec{y}}{|\vec{y}|}\right).$$

Although the scalar product is faster to compute, it enforces spherical clusters. Therefore we rely on the Mahalanobis distance in our approach.

3.2 Index Term Selection

To reduce the number of words in the vector description we applied a simple method for keyword selection by extracting keywords based on their entropy. In the approach discussed in [16], for each word k in the vocabulary the entropy as defined by [20] was computed:

$$W_k = 1 + \frac{1}{\log_2 n} \sum_{j=1}^{n} p_{jk} \log_2 p_{jk} \tag{12}$$

$$\text{with} \quad p_{jk} = \frac{\mathrm{tf}_{jk}}{\sum_{l=1}^{n} \mathrm{tf}_{lk}},$$

where tf_{jk} is the frequency of word k in document j, and n is the number of documents in the collection. Here the entropy gives a measure how well a word is suited to separate documents by keyword search. For instance, words that occur in many documents will have low entropy. The entropy can be seen as a measure of the importance of a word in the given domain context. As index words a number of words that have a high entropy relative to their overall frequency have been chosen, i.e. of words occurring equally often those with the higher entropy can be preferred. Empirically this procedure has been found to yield a set of relevant words that are suited to serve as index terms [16].

In order to obtain a fixed number of index terms that appropriately cover the documents, we applied a greedy strategy: From the first document in the set of documents select the term with the highest relative entropy as an index term. Then mark this document and all other documents containing this term. From the first of the remaining unmarked documents select again the term with the highest relative entropy as an index term. Then mark again this document and all other documents containing this term. Repeat this process until all documents are marked, then unmark all of them and start again. The process can be terminated when the desired number of index terms have been selected.

Label	Dataset Category	Associated Theme
A	Commercial Banks	Banking & Finance
B	Building Societies	Banking & Finance
C	Insurance Agencies	Banking & Finance
D	Java	Programming Lang.
E	C / C++	Programming Lang.
F	Visual Basic	Programming Lang.
G	Astronomy	Science
H	Biology	Science
I	Soccer	Sport
J	Motor Racing	Sport
K	Sport	Sport

Table 1: Categories and Themes of the used benchmark data set of web pages.

4 Experiments

For our experimental studies we chose the collection of web page documents used in [26].[1] The data set consists of 11,000 web pages classified into 11 equally-sized categories each containing 1,000 web documents. To each category one of four distinct themes, namely Banking and Finance, Programming Languages, Science, and Sport was assigned as shown in Table 1.

In the following we present results we obtained using the preprocessing strategies described above. After stemming and stop word filtering we obtained 163,860 words. This set was further reduced by removing terms that are shorter than 4 characters and that occur less then 15 times or more than $11,000/12 \approx 917$ times in the whole collection. In this way we made sure that no words that perfectly separate one class from another are used in the describing vectors. From the remaining 10626 words we selected 400 words by applying the greedy index-term selection approach described in Section 3.2. For our clustering experiments we selected finally subsets of the 20, 50, 100, 150, ..., 350, 400 most frequent words in the subset to be clustered. Based on these words we determined vector space descriptions for each document (see Section 3.1, Equation (10)) that we used in our clustering experiments. All vectors were normalized to unit length (*after* the subset selection).

To assess the clustering performance using term reweighting techniques, we computed the performance on the same data sets used in our previous experiments [5], i.e., we clustered the union of the dissimilar data sets A and I, and the semantically more similar data sets B and C. In a third experiment we used all classes and tried to find clusters describing the four main themes, i.e., banking, programming languages, science, and sport.

For our experiments we used c-means, fuzzy clustering and learning vector quantization methods. The learning

[1] This collection is available for download at http://www.pedal.rdg.ac.uk/banksearchdataset

vector quantization algorithm updated the cluster parameters once for every 100 documents.[2]

A detailed discussion of the performance of these methods with and without cluster centers normalized to unit length, with and without variances (i.e., spherical clusters and axes-parallel ellipsoids—diagonal covariance matrices—of equal size), and with the inverse squared distance or the Gaussian function for the activation can be found in [5]. Here, however, we focus on term re-weighting aspects. In order to evaluate the effects of keyword weighting, we computed the "importance" of each keyword for the classification of a document to a specific class based on different information theoretical measures. The "importance" values are then used to re-weight the terms in each document before the clustering process is started.

4.1 Keyword Weighting by Information Gain

Information gain (also known as *mutual (Shannon) information* or *(Shannon) cross entropy*), which is frequently used in decision tree learning, measures the average or expected entropy reduction resulting from finding out the value of a specific attribute. In text categorization information gain can be used to measure how well a term can be used to categorize a document, i.e., it measures the entropy reduction based on this specific term.

The information gain of a term t_k for a given set of r classes c_i is defined as:

$$\begin{aligned} I_{\text{gain}}(t_k) \quad = \quad & -\sum_{i=1}^{r} P(c_i) \log_2 P(c_i) \\ & +P(t_k) \sum_{i=1}^{r} P(c_i|t_k) \log_2 P(c_i|t_k) \\ & +P(\overline{t}_k) \sum_{i=1}^{r} P(c_i|\overline{t}_k) \log_2 P(c_i|\overline{t}_k) \end{aligned} \tag{13}$$

The information gain values are then used to re-weight the terms of each document by computing

$$x^*_{jk} \quad = \quad x_{jk} \cdot \max\{I_{\text{gain}}(t_k), 10^{-6}\}. \tag{14}$$

The results, obtained with cluster centers normalized to length 1 and no (co)variances, are shown in Figures 1 to 3. All results are the mean values of ten runs, which differed in the initial cluster positions and the order in which documents were processed. The dotted lines show the default accuracy (obtained if all documents are assigned to the majority class). The diamonds show the average classification accuracy in percent (left axis) with bars indicating the standard deviation. The grey dots and lines show the average execution times in seconds (right axis).

Re-weighting improved the performance for both two class problems (see Figures 1 and 2). Especially for the semantically similar data sets B and C a higher accurary is achieved and the clustering process is much more stable for fuzzy c-means and learning vector quantization (see the standard deviations in Figure 2). However, for the four class problem (Figure 3) no visible gains resulted. Rather, there is a slightly higher variation in the results of fuzzy c-means and learning vector quantization (see the standard deviations in Figure 3, middle row compared to top row).

[2] All experiments were carried out with a program written in C and compiled with gcc 3.3.3 on a Pentium 4C 2.6GHz system with 1GB of main memory running S.u.S.E. Linux 9.1. The program and its sources can be downloaded free of charge at http://fuzzy.cs.uni-magdeburg.de/~borgelt/cluster.html.

4.2 Keyword Weighting Using χ^2

The χ^2 measure, which is well known in statistics, uses the difference of expected and observed occurrences of attributes (here: terms) to measure the statistical significance of the attribute (here: goodness of a term) with respect to another attribute (here: the class). The χ^2 measure can be computed based on the contingency table of a term t and the classes c_i. If N is the total number of documents and r the number of classes, it can be defined as

$$\begin{aligned} \chi^2(t_k) \quad = \quad & N \sum_{i=1}^{r} \left(\frac{(P(c_i)P(t_k) - P(t_k, c_i))^2}{P(c_i)P(t_k)} \right. \\ & \left. + \frac{(P(c_i)P(\overline{t_k}) - P(\overline{t_k}, c_i))^2}{P(c_i)P(\overline{t_k})} \right). \end{aligned} \tag{15}$$

The resulting $\chi^2(t_k)$ values we used again to re-weight the terms, i.e., similar to Equation (14) we computed

$$x^*_{jk} \quad = \quad x_{jk} \cdot \max\{\chi^2(t_k), 10^{-6}\}. \tag{16}$$

Since the results using the χ^2 measure are almost identical to the results obtained with information gain, we do not show any diagrams. As it seems, both measures are equally well suited for term re-weighting.

4.3 Keyword Selection by Information Gain

Since re-weighting the terms based on information gain and the χ^2 measure did not improve the clustering performance in the four class case, we re-executed the experiments with 20, 50, and 100 terms, this time selecting them based on their rank w.r.t. information gain (in all other experiments the terms were chosen based on their frequency in the considered document set). The goal is to find out whether the weighting did not have any effect, because the wrong words are in the selected subset. The results are shown in the bottom row of Figure 3 (for higher numbers of values the results get identical to the ones shown in the middle row). Obviously, the classification performance was improved for all clustering methods, indicating that weighting alone is not sufficient if the terms are selected based on simple occurrence frequencies.

4.4 Keyword Extraction for Cluster Labeling

In order to obtain describing terms for each cluster, we compared different keyword rankings for clustering experiments for the four major themes using 200 terms. We selected a typical result for this discussion.

In Table 2 the terms are sorted descendingly by their information gain $I_{\text{gain}}(t_k)$ w.r.t. the found clusters (second column). Columns 3 to 6, each of which corresponds to a cluster, show the probability that term t_k appears in the corresponding cluster, with the highest probability per row set in bold face. Going down a column and collecting the highlighted terms yields a ranked keyword list for the corresponding cluster. Using $\chi^2(t_k)$ for the ranking leads to very similar results and thus we omit a separate table.

As an alternative, Table 3 shows the keyword ranking obtained by selecting the terms with the highest weights (largest center coordinates) for each cluster prototype. As can be seen, the order of the keywords is very similar to that of Table 2 (the rank differences are fairly small, at least in the shown top part of the list). Thus, to describe a cluster the terms can simply be selected from an ordered list of the center coordinates of each cluster, which is, of course, much more efficient than the computation of an information measure. At least for this example, the descriptive quality seems to be very similar.

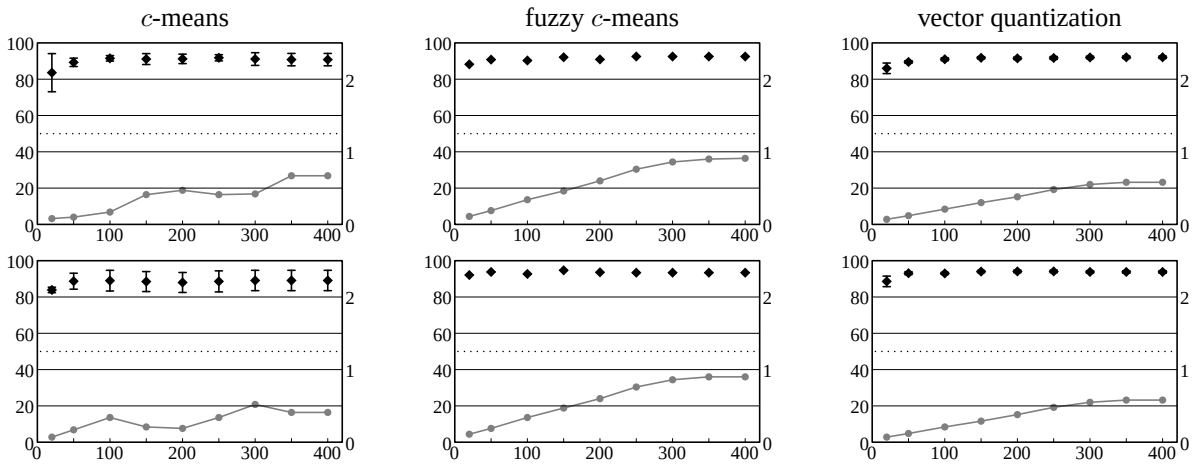

Figure 1: Accuracy on commercial banks versus soccer, normalized centers, fixed uniform variances (top row: no scaling, bottom row: scaling with information gain).

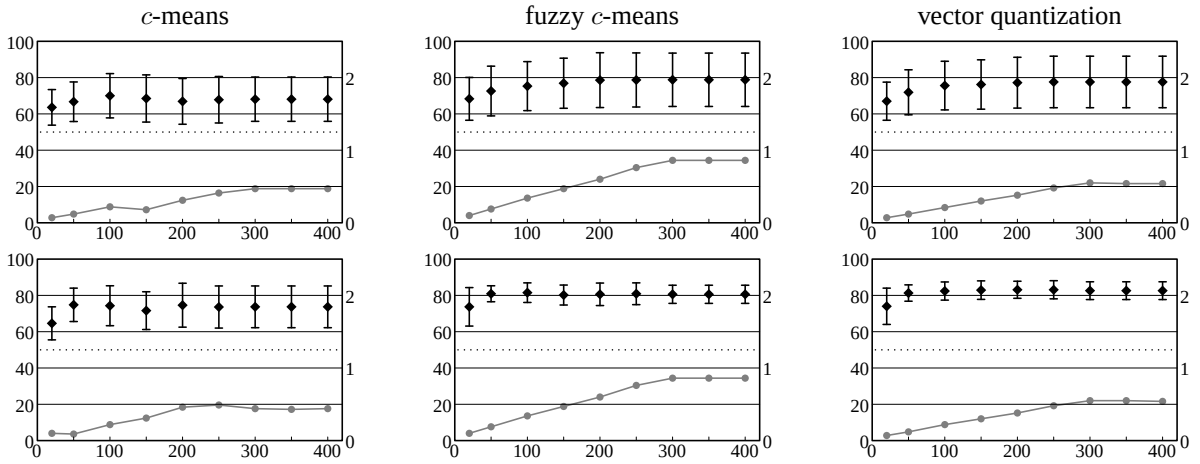

Figure 2: Accuracy on building companies versus insurance agencies, normalized centers, fixed uniform variances (top row: no scaling, bottom row: scaling with information gain).

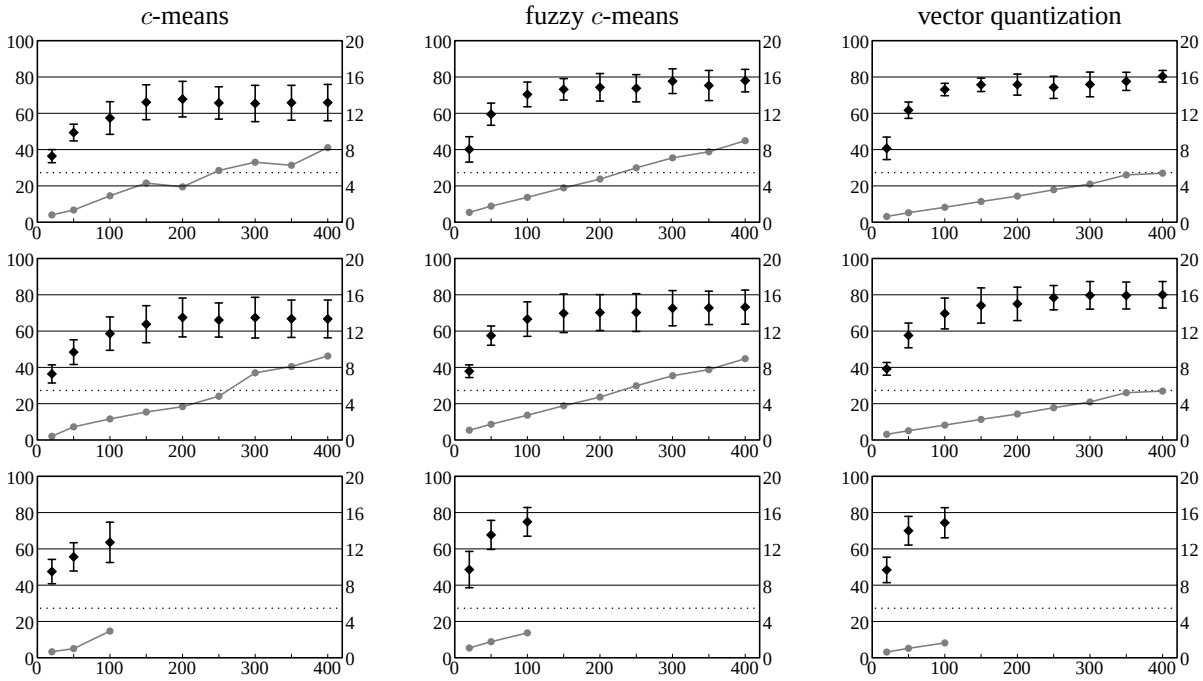

Figure 3: Accuracy on major themes (four clusters), normalized centers, fixed uniform variances (top row: no scaling, middle row: scaling with information gain, bottom row: no scaling, keyword selection based on information gain).

term t	$G(t)$	$p(t\|1)$	$p(t\|2)$	$p(t\|3)$	$p(t\|4)$
footbal	0.143	0.001	**0.299**	0.007	0.004
championship	0.120	0.001	**0.242**	0.001	0.002
galaxi	0.118	0.003	0.002	0.001	**0.270**
repay	0.112	0.004	0.000	**0.212**	0.000
genet	0.106	0.003	0.005	0.002	**0.253**
season	0.100	0.011	**0.277**	0.018	0.027
theori	0.094	0.017	0.012	0.004	**0.266**
applet	0.089	**0.178**	0.001	0.001	0.004
pension	0.087	0.004	0.004	**0.177**	0.001
astronom	0.084	0.002	0.003	0.000	**0.196**
energi	0.080	0.010	0.022	0.009	**0.249**
premium	0.080	0.006	0.006	**0.183**	0.008
button	0.080	**0.234**	0.018	0.030	0.011
sequenc	0.079	0.049	0.018	0.003	**0.254**
borrow	0.078	0.006	0.005	**0.180**	0.013
detect	0.076	0.073	0.005	0.005	**0.235**
telescop	0.076	0.003	0.003	0.003	**0.193**
cosmolog	0.075	0.003	0.002	0.000	**0.178**
coach	0.074	0.001	**0.173**	0.008	0.003
distanc	0.074	0.012	0.058	0.005	**0.239**
chequ	0.063	0.005	0.005	**0.140**	0.001
arrai	0.057	**0.148**	0.009	0.001	0.077
script	0.055	**0.150**	0.005	0.010	0.013
javascript	0.052	**0.147**	0.002	0.016	0.017
browser	0.051	**0.180**	0.011	0.025	0.050
gross	0.050	0.004	0.012	**0.134**	0.015
activex	0.050	**0.109**	0.001	0.002	0.002
liabil	0.050	0.005	0.006	**0.123**	0.006
surfac	0.050	0.020	0.028	0.004	**0.178**
default	0.049	**0.167**	0.009	0.030	0.020
input	0.049	**0.157**	0.016	0.008	0.053
purchas	0.048	0.035	0.035	**0.186**	0.019
⋮	⋮			⋮	

Table 2: Table of keywords ordered by information gain (second column). The values in columns 3 to 6 are the probabilities that the term appears in the given cluster.

applet	0.592	footbal	0.723
button	0.331	coach	0.329
javascript	0.319	championship	0.298
script	0.263	season	0.242
browser	0.188	david	0.132
password	0.153	basketbal	0.128
thread	0.150	round	0.122
arrai	0.148	scotland	0.100
⋮	⋮	⋮	⋮

repay	0.495	galaxi	0.626
borrow	0.324	genet	0.436
discount	0.312	sequenc	0.342
pension	0.281	theori	0.181
gross	0.226	distanc	0.171
chequ	0.225	cosmolog	0.170
premium	0.222	telescop	0.164
purchas	0.194	energi	0.156
⋮	⋮	⋮	⋮

Table 3: Tables of keywords for each cluster obtained by sorting the terms of each cluster prototype by their weight (center coordinate).

5 Conclusions

Our experiments show that prior information about the "importance" or "goodness" of a keyword for a desired class or cluster can improve the clustering performance for the separation of two classes. Since this information is frequently available in user profiles or by analyzing manually categorized document collections, "better" clusters can be obtained in this way.

For more than two classes, however, our results did not show a significant improvement. It would be interesting to analyze if a class specific computation of the information gain can improve the results. In such an approach the terms would have to be re-weighted individually for each cluster during the similarity computation. However, in this case the corresponding cluster centers also have to be initialized accordingly.

References

[1] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, NY, USA 1981

[2] J.C. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer, Dordrecht, Netherlands 1999

[3] J. Bilmes. A Gentle Tutorial on the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. University of Berkeley, Tech. Rep. ICSI-TR-97-021, 1997

[4] H.H. Bock. *Automatische Klassifikation*. Vandenhoeck & Ruprecht, Göttingen, Germany 1974

[5] C. Borgelt and A. Nürnberger. Fast Fuzzy Clustering of Web Page Collections. *Proc. PKDD Workshop on Statistical Approaches for Web Mining (SAWM, Pisa, Italy)*. 2004 (to appear).

[6] A.P. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society (Series B)* 39:1–38. Blackwell, Oxford, United Kingdom 1977

[7] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. J. Wiley & Sons, New York, NY, USA 1973

[8] B.S. Everitt and D.J. Hand. *Finite Mixture Distributions*. Chapman & Hall, London, UK 1981

[9] G. Forman. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research* 3:1289-1305, 2003

[10] H. Frigui and O. Nasraoui. Simultaneaous Clustering and Dynamic Keyword Weighting for Text Douments. M.W. Berry, ed. *Survey of Text Mining*, 45–72. 2003

[11] I. Gath and A.B. Geva. Unsupervised Optimal Fuzzy Clustering. *IEEE Trans. Pattern Analysis & Machine Intelligence* 11:773–781. IEEE Press, Piscataway, NJ, USA, 1989

[12] W.R. Greiff. A Theory of Term Weighting Based on Exploratory Data Analysis. *Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (Sydney, Australia)*, 17–19. ACM Press, New York, NY, USA 1998

[13] E.E. Gustafson and W.C. Kessel. Fuzzy Clustering with a Fuzzy Covariance Matrix. *Proc. 18th IEEE Conference on Decision and Control (IEEE CDC, San Diego, CA)*, 761–766, IEEE Press, Piscataway, NJ, USA 1979

[14] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis*. J. Wiley & Sons, Chichester, England 1999

[15] F. Klawonn and R. Kruse. Constructing a Fuzzy Controller from Data. *Fuzzy Sets and Systems* 85:177-193. North-Holland, Amsterdam, Netherlands 1997

[16] A. Klose, A. Nürnberger, R. Kruse, G.K. Hartmann, and M. Richards. Interactive Text Retrieval Based on Document Similarities. *Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy* 25:649–654. Elsevier, Amsterdam, Netherlands 2000

[17] T. Kohonen. *Learning Vector Quantization for Pattern Recognition*. Technical Report TKK-F-A601. Helsinki University of Technology, Finland 1986

[18] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Heidelberg, Germany 1995 (3rd ext. edition 2001)

[19] R. Krishnapuram and J. Keller. A Possibilistic Approach to Clustering, *IEEE Transactions on Fuzzy Systems*, 1:98-110. IEEE Press, Piscataway, NJ, USA 1993

[20] K.E. Lochbaum and L.A. Streeter. Combining and Comparing the Effectiveness of Latent Semantic Indexing and the Ordinary Vector Space Model for Information Retrieval. *Information Processing and Management* 25:665–676. Elsevier, Amsterdam, Netherlands 1989

[21] D. Mladenic. Using Text Learning to help Web browsing. *Proc. 9th Int. Conf. on Human-Computer Interaction*. New Orleans, LA, USA 2001

[22] M. Porter. An Algorithm for Suffix Stripping. *Program: Electronic Library & Information Systems* 14(3):130–137. Emerald, Bradford, United Kingdom 1980

[23] G. Salton, A. Wong, and C.S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM* 18:613–620 ACM Press, New York, NY, USA 1975

[24] G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management* 24:513–523. Elsevier, Amsterdam, Netherlands 1988

[25] G. Salton, J. Allan, and C. Buckley. Automatic Structuring and Retrieval of Large Text Files. *Communications of the ACM* 37:97–108. ACM Press, New York, NY, USA 1994

[26] M.P. Sinka, and D.W. Corne. A Large Benchmark Dataset for Web Document Clustering. *A. Abraham, J. Ruiz-del-Solar, and M. Köppen (eds.), Soft Computing Systems: Design, Management and Applications*, 881–890. IOS Press, Amsterdam, The Netherlands 2002

[27] H. Timm, C. Borgelt, and R. Kruse. A Modification to Improve Possibilistic Cluster Analysis. *Proc. IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2002, Honolulu, Hawaii)*. IEEE Press, Piscataway, NJ, USA 2002

[28] Y. Yang and J.O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. *Proc. 14th Int. Conf. on Machine Learning (ICML'97)*, 412–420. 1997