3rd Int. Conference on Artificial Intelligence and Applications (AIA 03).
Benalmádena, Spain, September 2003. Edited by M. H. Hanza.
pp. 216-221, ISBN 0-88986-390-3, ACTA Press, 2003.

On Cluster Validity and the Information Need of Users

Benno Stein Sven Meyer zu Eissen Frank Wißbrock

Paderborn University
Department of Computer Science
email: {stein, smze}@upb.de

Abstract

In the field of information retrieval, clustering algorithms are used to analyze large collections of documents with the objective to form groups of similar documents. Clustering a document collection is an ambiguous task: A clustering, i. e. a set of document groups, depends on the chosen clustering algorithm as well as on the algorithm's parameter settings. To find the best among several clusterings, it is common practice to evaluate their internal structures with a cluster validity measure.

A clustering is considered to be useful to a user if particular structural properties are well developed. Nevertheless, the presence of certain structural properties may not guarantee usefulness from an information retrieval standpoint, say, whether or not the found document groups resemble the classification of a human editor. The paper in hand investigates this point: Based on already classified document collections we generate clusterings and compare the predicted quality to their real quality.

Our analysis includes the classical cluster validity measures from Dunn and Davies-Bouldin as well as the new graph-based measures Λ (weighted edge connectivity) and $\overline{\rho}$ (expected edge density). The experiments show interesting results: The classical measures behave in a consistent manner insofar as mediocre and poor clusterings are identified as such. On real-world document clustering data, however, they are definitely outperformed by the expected edge density $\overline{\rho}$. This superiority of the graph-based measures can be explained by their independence of cluster forms and distances.

Key words
Data Mining, Document Categorization, Cluster Validity Measures, Cluster Usability

1 Cluster Validity versus Cluster Usability

Clustering is a useful method to analyze large collections of documents. It has the potential to identify unknown classification schemes that highlight relations and differences between documents. There is a large number of clustering algorithms that can be used for this task. But different clustering algorithms tend to produce different results and even a single clustering algorithm creates various results depending on its initial parameter settings. Therefore an evaluation of the results is necessary to assess their quality. In clustering tasks the procedure of evaluating the results is known under the term cluster validity [1].

Most cluster validity measures assess certain structural properties of a clustering result. If the structural properties of the outcome are well developed, then the result is considered to be of interest to the user. If the structural properties are not well developed, then the result is considered to be of no interest. Because the focus is on the structural properties of a data set, these measures are also called objective measures [2]. Many objective measures have been developed so far; their attractiveness stems from their domain independence. Every clustering result, regardless which algorithm produced it, can be evaluated whether it has certain structural properties or not.

High scores on an objective measure make a clustering result valid with regard to its structural properties—however, the presence of such structural properties does not guarantee the interestingness of the result for the user: Objective measures lack the linkage to the user's information need. Recall that in the field of pattern recognition, measures that consider a user's information need are referred to as subjective measures [2]. To find a corresponding term in the field of clustering result evaluation, one could speak of *cluster usability*. Research on subjective measures has not been as intensive as on objective measures, and there are not many texts that discuss subjective measures in the context of document clustering [3].

The goal of our research is to identify subjective measures for document clustering tasks. This is not only of academic concern to us: We have developed the meta search engine AISEARCH [4], and we want to improve the classification of the delivered search results.

In this paper we embark on the following strategy: We investigate up to what extent certain cluster validity measures can be employed to predict cluster usability. Note that for a carefully classified document collection D the usability of a clustering $\mathcal{C}$ of D can be quantified easily, e. g. by means of the F-Measure, which evaluates the quality of the best match between the original classes and the found clusters. In particular, we want to answer the following question:

> *Which of the investigated cluster validity measures qualifies a user's information need, say, captures structural properties that correlate with the F-Measure?*

We answer this question by performing various clustering experiments based on the new Reuters Text Corpus Volume 1 (RCV1) [5]. In this connection we employ the following clustering algorithms: k-Means (iterative), Group-Average-Link (hierarchical agglomerative), and MAJORCLUST (density-based) [6, 7, 8, 9, 10]. Note that the generation of the clusterings as well as the cluster algorithms are not discussed here. The remainder of this paper is organized as follows. Section 2 introduces the investigated cluster validity measures, and Section 3 presents the analysis results.

2 Cluster Validity Measures

Definition 1 (Clustering) *Let D be a set of objects. A clustering $\mathcal{C} = \{C \mid C \subseteq D\}$ of D is a division of D into sets for which the following conditions hold: $\bigcup_{C_i \in \mathcal{C}} C_i = D$, and $\forall C_i, C_j \in \mathcal{C} : C_i \cap C_{j \neq i} = \emptyset$. The sets C_i are called clusters.*

Here, the set of objects, D, corresponds to a document collection. Moreover, it is useful to consider the elements in D as nodes of a weighted graph G. G is completely connected, and the weight of the edge that connects two documents, d_i, d_j, corresponds to their similarity.

A cluster validity measure maps a clustering on a real number. The number indicates to what degree certain structural properties are developed in the clustering. There are two types of measures: external measures and internal measures. The external measures use a human reference classification to evaluate the clustering. Note that external measures are not applicable in real world situations since reference classifications are usually not available. In contrast, internal measures base their calculations solely on the clustering that has to be evaluated. In the follwing, five cluster validity measures are described. One of them, the F-Measure, is an external measure and was used in our experiments to evaluate the other measures. The remaining four measures are internal: Dunn and Davies-Bouldin, which are widely accepted classical measures, and Λ and $\overline{\rho}$, which were developed at our Institute [10].

2.1 The F-Measure

The F-Measure combines the precision and recall measures from information retrieval [11].

Definition 2 (Precision, Recall, F-Measure) *Let D represent the set of documents and let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a clustering of D. Moreover, let $\mathcal{C}^* = \{C_1^*, \ldots, C_l^*\}$ designate the human reference classification.*

Then the recall of cluster j with respect to class i, $rec(i,j)$, is defined as $|C_j \cap C_i^*|/|C_i^*|$. The precision of cluster j with respect to class i, $prec(i,j)$, is defined as $|C_j \cap C_i^*|/|C_j|$. The F-Measure combines both values as follows:

$$F_{i,j} = \frac{2}{\frac{1}{prec(i,j)} + \frac{1}{rec(i,j)}}$$

Based on this formula, the overall F-Measure of a clustering is:

$$F = \sum_{i=1}^{l} \frac{|C_i^*|}{|D|} \cdot \max_{j=1,\ldots,k} \{F_{i,j}\}$$

Note that a perfect fit between clustering and human reference classification leads to a F-Measure score of 1, which is the maximal possible value of the measure.

2.2 The Dunn Index Family

Definition 3 (Dunn Measure) *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a clustering of a set of objects D, $\delta : \mathcal{C} \times \mathcal{C} \rightarrow \mathbf{R}$ be a cluster to cluster distance measure, and $\Delta : \mathcal{C} \rightarrow \mathbf{R}$ be a cluster diameter measure. Then all measures I of the form*

$$I(\mathcal{C}) = \frac{\min_{i \neq j}\{\delta(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}}$$

are called Dunn indices.

Originally, Dunn used

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x,y) \quad \text{and}$$

$$\Delta(C_i) = \max_{x,y \in C_i} d(x,y)$$

where $d : D \times D \rightarrow \mathbf{R}$ is a function that measures the distance between objects of D. With these settings the measure yields high values for clusterings with compact and very well separated clusters. The upper part of Figure 1 gives an example. The maximum diameter is very low and the minimum distance between two clusters is relatively large. As a consequence, the clustering is ranked high by the Dunn index. However, Bezdek recognized that the index is very noise sensitive (see Figure 1 bottom): Even though the clustering in the example is good, the large diameter of C_1 along with the small distance between C_3 and C_4 leads to a low value of the Dunn index. Bezdek experienced that the combination of

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x,y) \quad \text{and}$$

$$\Delta(C_i) = 2\left(\frac{\sum_{x \in C_i} d(x, c_i)}{|C_i|}\right)$$

gave reliable results for several data sets from different domains [12]. Here, c_i denotes the centroid of cluster C_i. Note that a maximization of I is desired.

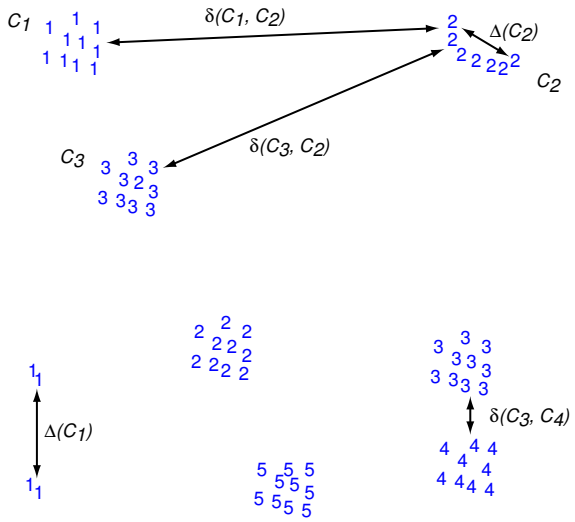

Figure 1. The upper part shows a clustering with compact and well separated clusters. The maximum diameter is low and the distance of the two closest clusters is large, and the original Dunn index returns a high score for the clustering. The clustering in the lower part contains some noise. The large diameter of the left cluster and the closeness of the two right clusters are considered to be typical for the clustering. As a consequence, the index returns a low value even though the clustering fits the strucure well.

2.3 The Davies-Bouldin Index

Davies and Bouldin proposed the following index that is known as Davies-Bouldin measure [13]. It is a function of the ratio of the sum of within-cluster scatter to between-cluster separation.

Definition 4 (Davies-Bouldin Index) *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a clustering of a set D of objects.*

$$DB = \frac{1}{k} \cdot \sum_{i=1}^{k} R_i, \quad \textit{with}$$

$$R_i = \max_{\substack{j=1,\ldots,n,\\ i \neq j}} R_{ij} \quad \textit{and} \quad R_{ij} = \frac{(s(C_i) + s(C_j))}{\delta(C_i, C_j)},$$

where $s : \mathcal{C} \to \mathbf{R}$ measures the scatter within a cluster, and $\delta : \mathcal{C} \times \mathcal{C} \to \mathbf{R}$ is a cluster to cluster distance measure.

Given the centroids c_i of the clusters C_i, a typical scatter measure is $s(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} ||x - c_i||$, and a typical cluster to cluster distance measure is the distance between the centroids, $||c_i - c_j||$. Because a low scatter and a high distance between clusters lead to low values of R_{ij}, a minimization of DB is desired.

Remarks. The Dunn index and the Davies-Bouldin index are related in that they have a geometric (typically centroidic) view on the clustering. The measures work well if the underlying data contains clusters of spherical form, but they are susceptible to data where this condition does not hold. Λ as well as $\overline{\rho}$ interpret a data set as a weighted similarity graph; they analyze the graph's edge density distribution to judge the quality of a clustering. Both measures are introduced in the following.

2.4 The Λ-Measure

A document collection can be considered as a weighted graph $G = \langle V, E, w \rangle$ with node set V, edge set E, and weight function $w : E \to [0, 1]$ where V represents the documents, and w defines the similarities between two adjacent documents. The Λ measure computes the weighted partial connectivity of $G = \langle V, E, w \rangle$, which is defined below. Observe that higher values of Λ indicate a better clustering.

Definition 5 (weighted partial connectivity Λ) *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a clustering of the nodes V of a weighted graph $G = \langle V, E, w \rangle$.*

$$\Lambda(\mathcal{C}) := \sum_{i=1}^{k} |C_i| \cdot \lambda_i,$$

where λ_i designates the weighted edge connectivity of $G(C_i)$. The weighted edge connectivity, λ, of a graph $G = \langle V, E, w \rangle$ is defined as $\min \sum_{\{u,v\} \in E'} w(u, v)$ where $E' \subset E$ and $G' = \langle V, E \setminus E' \rangle$ is not connected. λ is also designated as the capacity of a minimum cut of G.

2.5 A Measure of Expected Density: $\overline{\rho}$

A graph $G = \langle V, E, w \rangle$ is called sparse if $|E| = \mathcal{O}(|V|)$; it is called dense if $|E| = \mathcal{O}(|V|^2)$. Put another way, we can compute the density θ of a graph from the equation $|E| = |V|^\theta$. With $w(G) := |V| + \sum_{e \in E} w(e)$, this relation extends naturally to weighted graphs:[1]

$$w(G) = |V|^\theta \quad \Leftrightarrow \quad \theta = \frac{\ln\big(w(G)\big)}{\ln\big(|V|\big)}$$

Obviously, θ can be used to compare the density of each induced subgraph $G' = \langle V', E', w' \rangle$ of G to the density of G: G' is sparse (dense) compared to G if the quotient $w(G')/(|V'|^\theta)$ is smaller (larger) than 1. This consideration is the key idea behind the following definition of a clustering's expected density $\overline{\rho}$.

Definition 6 (expected density $\overline{\rho}$) *Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a clustering of a weighted graph $G = \langle V, E, w \rangle$, and let $G_i = \langle V_i, E_i, w_i \rangle$ be the induced subgraph of G with respect to cluster C_i. Then the expected density of a clustering $\mathcal{C}$ is defined as follows.*

$$\overline{\rho}(\mathcal{C}) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^\theta}, \quad \textit{where} \quad |V|^\theta = w(G)$$

Since the edge weights resemble the similarity of the objects which are represented by V, a higher value of $\overline{\rho}$ indicates a better clustering.

[1] $w(G)$ denotes the total edge weight of G plus the number of nodes, $|V|$, which serves as adjustment term for small graphs.

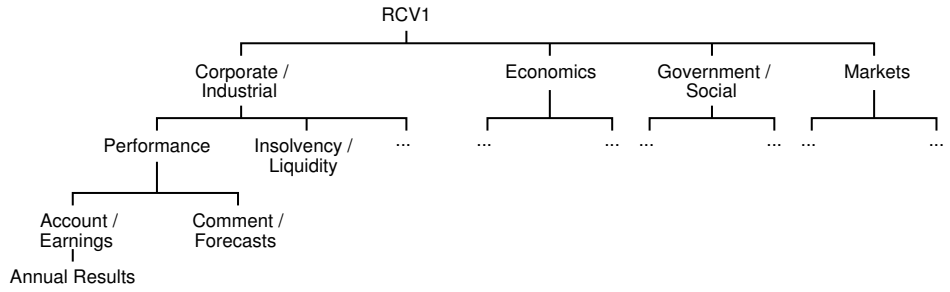

Figure 2. Part of the RVC1 category structure. There are four top-level categories. Every node under the top-level categories represent a specialization of its parent.

3 Test Environment and Test Settings

The experiments have been conducted with samples of RCV1, short hand for "Reuters Corpus Volume 1" [5]. RCV1 is a document collection that was published by the Reuters Corporation for research purposes. It contains over 800,000 documents each of which consisting of a few hundred up to several thousands words. The documents are enriched by meta information like category (also called topic), geographic region, or industry sector. There are 103 different categories, which are arranged within a hierarchy of the four top level categories "Corporate/Industrial", "Economics", "Government/Social", and "Markets". Each of the top level categories defines the root of a tree of subcategories, where every child node fine grains the information given by its parent (cf. Figure 2). Note that a document d can be assigned to several categories $c_1, \ldots, c_p$, and that all ancestor categories of a category c_i are assigned to d as well.

For our experiments, we considered two documents d_1, d_2 as belonging to the same category c_s if they share both the same top level category c_t and the same most specific category c_s. Moreover, we constructed the test sets in such a way that there is no document d_1 whose most specific category c_s is an ancestor of the most specific category of some other document d_2.

The number of categories in our test data varies from three to six. For each category, between 100 and 300 documents were drawn randomly from the entire category. The data sets had different sizes and class numbers; we investigated uniformly as well as non-uniformly distributed category sizes. Table 1 gives an overview of the constructed data sets.

The preprocessing of the documents included parsing of text body and title, stop word removal according to standard stop word lists, the application of Porter's stemming algorithm [14], and indexing according to term frequency. We used the standard cosine similarity measure to capture the similarities between documents.

Three analyses were conducted on each test data set to evaluate the performance of the aforementioned indices. The three analyses along with their results are described in the next three subsections.

	DS1	DS2	DS3	DS4	DS5	DS6	DS7
# categories	3	3	4	3	5	5	6
# documents	300	600	400	450	500	800	900
unif. distributed	yes	yes	yes	yes	yes	no	no

Table 1. Overview of the constructed data sets.

3.1 Consistence Analysis

Since we know the reference categorization $\mathcal{C}^*$ which was provided by a human editor, we can use it to generate artificial clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_n$ that are to a greater or lesser extent modifications of C^*. The F-Measure values for $\mathcal{C}_1, \ldots, \mathcal{C}_n$ will measure the degree of congruence for the modified sets with respect to $\mathcal{C}^*$. Assuming that the modified categorizations represent erroneous clusterings, the value of a validation index for $\mathcal{C}_1, \ldots, \mathcal{C}_n$ should be worse than for $\mathcal{C}^*$. Even more can be expected: For $\mathcal{C}_1, \ldots, \mathcal{C}_n$, the values of a subjective validation index should relate to the values of the F-Measure monotonically.

To derive an artificial clustering $\mathcal{C}_i$ of $\mathcal{C}^*$, we repeatedly chose two distinct clusters of $\mathcal{C}^*$ and interchanged randomly chosen subsets of documents pairwise between the clusters. Note that the size of the interchanged subsets controls the degree of congruence between $\mathcal{C}_i$ and $\mathcal{C}^*$. We varied the sizes of the subsets between 1 document and 50% of the documents within a cluster.

The Figures 3, 4 and 5 show the resulting scatter plots for the artificial clusterings that we derived from the reference categorization of DS6 and evaluated with the data set DS6. We measured the F-Measure value (y-axis) and the validity index value (x-axis) for each clustering. For the sake of better readability, we changed the sign of the Davies-Bouldin Index, which is the only one to be minimized—this way, the plots are directly comparable. Assuming that a greater index value constitutes a better clustering, an ideal index would show points on a curve that starts in the lower left corner and grows monotonically up to the upper right corner. Observe that all of the indices perform well on this test on all data sets.

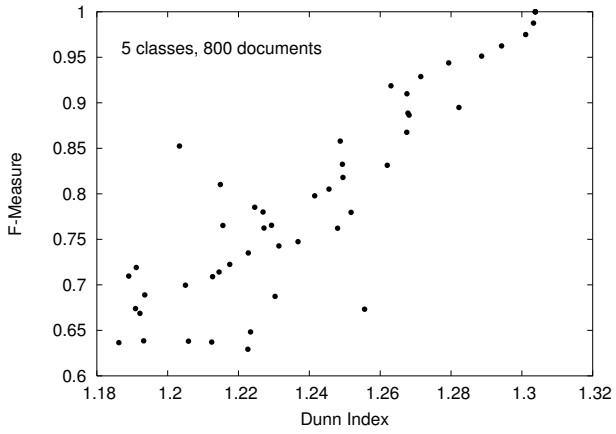

Figure 3. Correlation of the F-Measure and the Dunn Index for the artificial clusterings.

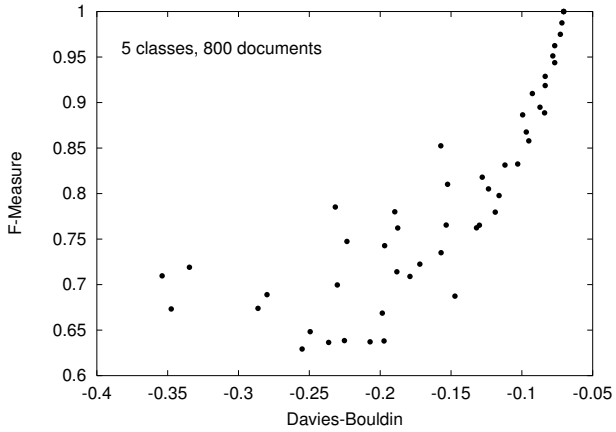

Figure 4. Correlation of the F-Measure and the Davies-Bouldin Index for the artificial clusterings.

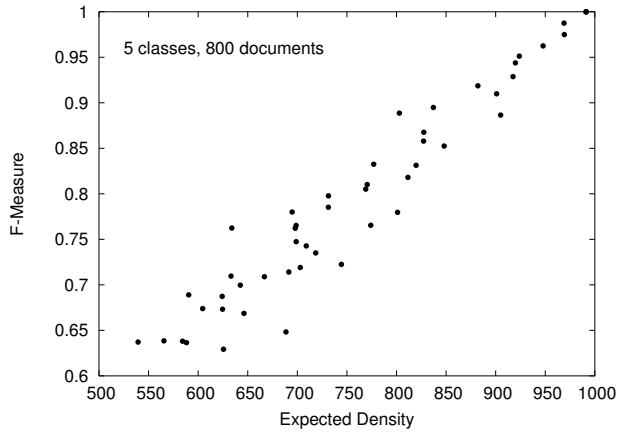

Figure 5. Correlation of the F-Measure and the $\overline{\rho}$-Measure for the artificial clusterings.

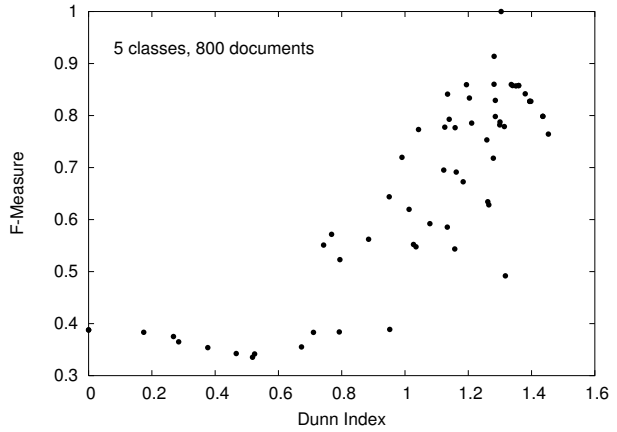

Figure 6. Correlation of the F-Measure and the Dunn Index for genuine clusterings.

3.2 Analysis with Genuine Clusterings

We are normally faced with clusterings which are not artificially constructed but stem from a document categorization system that uses different clustering algorithms. For the experiments reported below we employed hierarchical, iterative, and density-based algorithms. Moreover, for each of these algorithms different thresholds, agglomeration levels, cluster numbers, etc. were tried. We measured the F-Measure values and corresponding validity index value for each clustering, and, in particular, for the reference categorization $\mathcal{C}^*$ that can be identified by its F-Measure value of 1.

The Figures 6, 7, and 8 show representative scatter plots for the investigated validity indices for the same data set that was used for the consistency analysis (DS6). Again, we changed the sign of the Davies-Bouldin Index.

Note that the best Dunn Index values do not correlate with the best F-Measure values. Also note that the Davies-Bouldin Index, which works well for the synthetic data sets, gets misleaded by the genuine clusterings generated by our clustering algorithms: Many clusterings with low F-Measure values untruly obtain a high index value.

3.3 Prediction Quality

One might argue that a validity index only has to find the best clustering among several candidates—a single outlier which is characterized by a top index value but a poor clustering can completely ruin the applicability of the index. Therefore we measured the F-Measure value that corresponds to the maximum index value for each validity index and each data set. Table 2 comprises the results.

	DS1	DS2	DS3	DS4	DS5	DS6	DS7	average
Dunn	0.69	0.68	0.75	0.75	0.75	0.76	0.61	0.71
D.-B.	0.50	0.50	0.50	0.40	0.33	0.39	0.35	0.42
Λ	1.00	0.78	0.67	0.40	0.66	0.90	0.85	0.75
$\overline{\rho}$	0.87	0.78	0.98	0.97	0.69	1.00	0.77	0.87

Table 2. The table shows for each index the F-Measure values that belong to its top-rated clusterings. Since the reference categorization $\mathcal{C}^*$ was among the evaluated clusterings, a perfect prediction corresponds to the F-Measure value of 1.

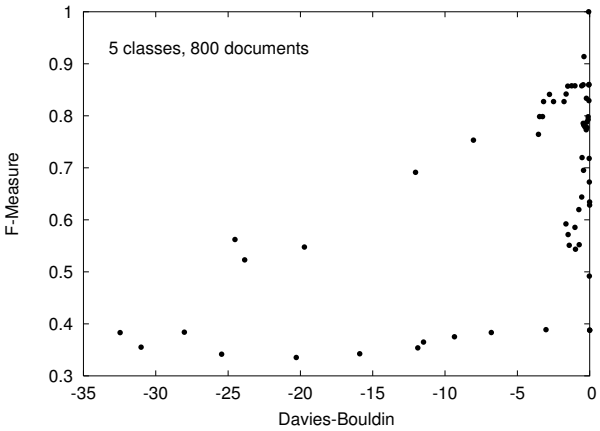

Figure 7. Correlation of the F-Measure and the Davies-Bouldin Index for genuine clusterings.

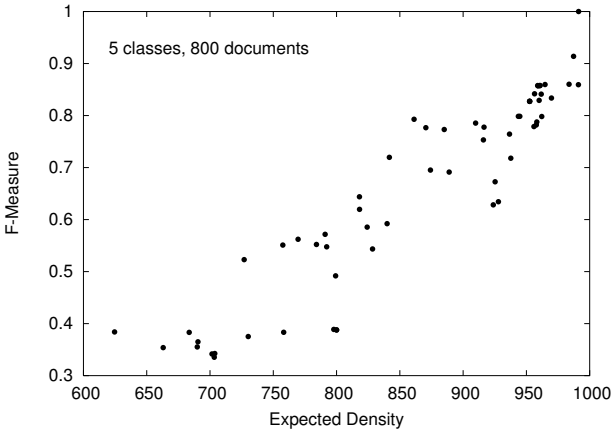

Figure 8. Correlation of the F-Measure and the $\overline{\rho}$-Measure for genuine clusterings.

Summary

Clustering algorithms are considered as a technology that has the potential to automatically analyze large collections of documents. Different clustering algorithms produce different clusterings, and cluster validity measures must be applied to identify among a set of clusterings the most valuable one. Cluster validity measures assess structural properties of a clustering—they hence are in the role of objective measures. The key question in this connection is whether or not an objective measure can be used to capture a user's information need.

In the field of automatic document categorization the information need corresponds to the categorization quality of a clustering $\mathcal{C}$. Given a reference categorization $\mathcal{C}^*$ the categorization quality of $\mathcal{C}$ can be quantified with the achieved precision and recall values. I. e., in the field of document categorization an answer to the above question can be given by analyzing the correlation of cluster validity measures with the F-Measure.

The paper in hand presents the related experiments. It investigates the classical cluster validity measures from Dunn and Davies-Bouldin and presents the new graph-based measures Λ and $\overline{\rho}$. As reported in the experiment section, the new $\overline{\rho}$-Measure performed convincingly on both artificial and genuine clusterings of different document sets, and it outperformed the classical measures in this domain. The Dunn Index performed robust but often missed to discover the real interesting clusterings. The Davies-Bouldin index performed well on artificial data sets—however, it was not able to correctly select the best clustering among clusterings that stemmed from a genuine document cluster application.

References

1. Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Clustering Validity Checking Methods: Part II. *ACM SIGMOD Record*, 31(3):19–27, 2002. ISSN 0163-5808.
2. A. Silberschatz and A. Tuzhilin. What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Trans. on Knowledge and Data Engineering*, 8(6), 1996.
3. A. Tuzhilin. *Handbook of Data Mining and Knowledge Discovery*, chapter Usefulness, Novelty, and Integration of Interestingness Measures. Oxford University Press, 2002.
4. Sven Meyer zu Eißen and Benno Stein. The AISEARCH Meta Search Engine Prototype. In *Proc. 12th Workshop on Information Technology and Systems (WITS 02), Barcelona Spain*. Technical University of Barcelona, Dec. 2002.
5. T.G. Rose, M. Stevenson, and M. Whitehead. The Reuters Corpus Volume 1 - From Yesterday's News to Tomorrow's Language Resources. In *Proc. 3rd International Conference on Language Resources and Evaluation*, 2002.
6. J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
7. Leonard Kaufman and Peter J. Rousseuw. *Finding Groups in Data*. Wiley, 1990.
8. Anil K. Jain and Richard C. Dubes. *Algorithm for Clustering in Data*. Prentice Hall, Englewood Cliffs, NJ, 1990. ISBN 0-13-022278-X.
9. B. S. Everitt. Cluster analysis. New York, Toronto, 1993.
10. Benno Stein and Oliver Niggemann. *25. Workshop on Graph Theory*, chapter On the Nature of Structure and its Identification. LNCS. Springer, Ascona, Italy, July 1999.
11. B. Larsen and Ch. Aone. Fast and Effective Text Mining Using Linear-time Document Clustering. In *Proc. KDD-99 Workshop San Diego USA*, San Diego, CA, USA, 1999.
12. J. C. Bezdek, W. Q. Li, Y. Attikiouzel, and M. Windham. A Geometric Approach to Cluster Validity for Normal Mixtures. *Soft Computing 1*, September 1997.
13. D.L. Davies and D.W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Learning*, 1(2), 1979.
14. M. F. Porter. An algorithm for suffix stripping. *Program*, 14 (3):130–137, 1980.