A new efficient and unbiased approach for clustering quality evaluation

Jean-Charles Lamirel, Pascal Cuxac, Raghvendra Mall

A new efficient and unbiased approach for clustering quality evaluation

Jean-Charles Lamirel[1], Pascal Cuxac[2], and Raghvendra Mall[3]

[1]LORIA, Campus Scientifique,
BP 239, Vandœuvre-lès-Nancy, France
jean-charles.lamirel@inria.fr
http://www.loria.fr,
[2]INIST-CNRS, 2 allée du Parc de Brabois,
54500 Vandœuvre-lès-Nancy, France
pascal.cuxac@inist.fr
http://recherche.inist.fr,
[3]Center of Data Engineering, IIIT Hyderabad,
NBH-61, Hyderabad, Andhra Pradesh, India
raghvendra.mall@research.iiit.ac.in
http://www.iiit.ac.in

Abstract. Traditional quality indexes (Inertia, DB, ...) are known to be method-dependent indexes that do not allow to properly estimate the quality of the clustering in several cases, as in that one of complex data, like textual data. We thus propose an alternative approach for clustering quality evaluation based on unsupervised measures of Recall, Precision and F-measure exploiting the descriptors of the data associated with the obtained clusters. Two categories of index are proposed, that are Macro and Micro indexes. This paper also focuses on the construction of a new cumulative Micro precision index that makes it possible to evaluate the overall quality of a clustering result while clearly distinguishing between homogeneous and heterogeneous, or degenerated results. The experimental comparison of the behavior of the classical indexes with our new approach is performed on a polythematic dataset of bibliographical references issued from the PASCAL database.

1 Introduction

The use of classification methods is mandatory for analyzing large corpus of data as it is the case in the domain of scientific survey or in that of strategic analyzes of research. While carrying out a classification, one seeks to build homogeneous groups of data sharing a certain number of identical characteristics. Furthermore, the clustering, or unsupervised classification, makes it possible to highlight these groups without prior knowledge on the treated data. A central problem that then arises is to qualify the obtained results in terms of quality: a quality index is a criterion which indeed makes it possible altogether to decide which clustering method to use, to fix an optimal number of clusters, and to evaluate or to develop a new method. Even if there exist recent alternative approaches [2] [8] [9], the

most usual indexes employed for the evaluation of the quality of clustering are mainly distance-based indexes relying on the concepts of intra cluster inertia and inter-cluster inertia [14]:

- Intra-cluster inertia measures the degree of homogeneity between the data associated with a cluster. It calculates their distances compared to the reference point representing the profile of the cluster. It can be defined as:

$$Intra = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|c|} \sum_{d \in c} \parallel p_c - p_d \parallel^2$$

  where C represents the set of clusters associated to the clustering result, d represents a cluster associated data and p_x represents the profile vector associated to the element x.

- Inter-clusters inertia measures the degree of heterogeneity between the clusters. It calculates the distances between the reference points representing the profiles of the various clusters of the partition.

$$Inter = \frac{1}{|C|^2 - |C|} \sum_{c \in C} \sum_{c' \in C, c' \neq c} \parallel p_c - p_{c'} \parallel^2$$

Thanks to these two quality indexes or their adaptations, like the Dunn index [5], the Davies-Bouldin index [3], or the Silhouette index [18], a clustering result is considered as good if it possesses low intra-clusters distances as compared to its inter-clusters distances. However, it has been shown in [12] that the distance based indexes are often strongly biased and highly dependent on the clustering method. They cannot thus be easily used for comparing different methods. Moreover, as Forest also pointed out [6], the experiments on these indexes in the literature are often performed on unrealistic test corpora constituted of low dimensional data and embedding a small number of potential classes. As an example, in their reference paper Milligan and Cooper [17] compared 30 different methods for estimating the number of clusters relying only on simulated data described in a low dimensional Euclidean space. Nonetheless, using Reuters test collection, it has been shown by Kassab and Lamirel [10] that aforementioned indexes are often properly unable to identify an optimal clustering model whenever the dataset is constituted by complex data that must be represented in a both highly multidimensional and sparse description space, as it is often the case with textual data. To cope with such problems, our own approach takes its inspiration both from the behavior of symbolic classifiers and from the evaluation principles used in Information Retrieval (IR). Our Recall/Precision and F-measures indexes exploit the properties of the data associated to each cluster after the clustering process without prior consideration of clusters profiles [12]. Their main advantage is thus to be independent of the clustering methods and of their operating mode. However, our last experiments highlighted that these new quality indexes did not make it possible to clearly distinguish between homogeneous results of clustering and heterogeneous, or degenerated ones

[7]. After presenting our original quality indexes, we thus describe hereafter some of their extensions which make it possible to solve the aforementioned problem. We then experimentally show the effectiveness of our extended approach, as compared to classical distance-based approach, for discriminating between the results provided by three different clustering methods which have been applied on a polythematic documentary corpus containing various bibliographic records issued from the PASCAL CNRS scientific database.

2 Unsupervised Recall Precision F-measure indexes

2.1 Overall clustering quality estimation

In IR, the **Recall R** represents the ratio between the number of relevant documents which have been returned by an IR system for a given query and the total number of relevant documents which should have been found in the documentary database [19]. The **Precision P** represents the ratio between the number of relevant documents which have been returned by an IR system for a given query and the total number of documents returned for the said query. **Recall** and **Precision** generally behave in an antagonist way: as **Recall** increases, **Precision** decreases, and conversely. The **F** function has thus been proposed by Van Rijsbergen [20] in order to highlight the best compromise between these two values. It is given by:

$$F = \frac{2(R * P)}{R + P} \tag{1}$$

Based on the same principles, the *Recall* and *Precision* indexes which we introduce hereafter evaluate the quality of a clustering method in an unsupervised way[1] by measuring the relevance of the clusters content in terms of shared properties, or features. In our further descriptions, a cluster content is supposed to be represented by the data associated with this latter after the clustering process and the descriptors (i.e. the properties or features) of the data are supposed to be weighted by values within the range [0,1].

Let us consider a set of clusters C resulting from a clustering method applied on a set of data D, the local *Recall* (Rec) and *Precision* (Prec) indexes for a given property p of the cluster c can be expressed as:

$$\underset{c}{Rec}(p) = \frac{|c_p^*|}{|D_p^*|}, \underset{c}{Prec}(p) = \frac{|c_p^*|}{|c|}$$

where the notation X_p^* represents the restriction of the set X to the set members having the property p.

[1] Conversely to classical Recall and Precision indexes that are supervised.

Then, for estimating the overall clustering quality, the averaged *Macro-Recall* (R_M) and *Macro-Precision* (P_M) indexes can be expressed as:

$$R_M = \frac{1}{|\overline{C}|}\sum_{c\in\overline{C}}\frac{1}{|S_C|}\sum_{p\in S_C}\underset{c}{Rec}(p), P_M = \frac{1}{|\overline{C}|}\sum_{c\in\overline{C}}\frac{1}{|S_C|}\sum_{p\in S_C}\underset{c}{Prec}(p) \quad (2)$$

where S_c is the set of peculiar properties of the cluster c, which can be defined as:

$$S_c = \left\{ p \in d, d \in c | \overline{W_c^p} = Max_{c'\in C}\left(\overline{W}_{c'}^p\right) \right\} \quad (3)$$

and where $\overline{C}$ represents the peculiar set of clusters extracted from the clusters of C, which verifies:

$$\overline{C} = \{c \in C | S_c \neq \emptyset\}$$

and, finally:

$$\overline{W}_c^p = \frac{\sum_{d\in c} W_d^p}{\sum_{c'\in C}\sum_{d\in c'} W_d^p} \quad (4)$$

where W_x^p represents the weight of the property p for element x.

It can be demonstrated (see [12] for more details) that if both values of averaged *Recall* and *Precision* reach the unity value, the peculiar set of clusters $\overline{C}$ represents a Galois lattice. Therefore, the combination of this two measures enables to evaluate to what extent a numerical clustering model can be assimilated to a Galois lattice natural classifier.

Macro-Recall and *Macro-Precision* indexes defined by (Eq. 2) can be considered as cluster-oriented measures because they provide average values of *Recall* and *Precision* for each cluster. They have opposite behaviors according to the number of clusters. Thus, these indexes permit to estimate in a global way an optimal number of clusters for a given method and a given dataset. The best data partition, or clustering result, is in this case the one which minimizes the difference between their values (see Figure 1B). However, similarly to the classical distance-based indexes, their main defect is that they do not permit to detect degenerated clustering results, whenever those jointly include a small number of heterogeneous or "garbage" clusters of large size and a big number of "chunk" clusters of very small size [7]. To correct that, we propose to construct complementary property-oriented indexes of *Micro-Recall* and *Micro-Precision* by averaging the *Recall/Precision* values of the peculiar properties independently of the structure of the clusters:

$$R_m = \frac{1}{|L|}\sum_{c\in\overline{C}, p\in S_c}\underset{c}{Rec}(p), P_m = \frac{1}{|L|}\sum_{c\in\overline{C}, p\in S_c}\underset{c}{Prec}(p) \quad (5)$$

where L represents the size of the data description space.

It is possible to refer not only to the information provided by the indices *Micro-Precision* and *Micro-Recall*, but to the calculation of the *Micro-Precision* operated cumulatively. In the latter case, the idea is to give a major influence to large clusters which are most likely to repatriate the heterogeneous information, and therefore, by themselves, lowering the quality of the resulting clustering. This calculation can be made as follows:

$$CP_m = \frac{\sum_{i=|c_{inf}|,|c_{sup}|} \frac{1}{|C_{i+}|^2} \sum_{c \in C_{i+}, p \in S_c} \frac{|c_p|}{|c|}}{\sum_{i=|c_{inf}|,|c_{sup}|} \frac{1}{C_{i+}}} \tag{6}$$

where C_{i+} represents the subset of clusters of C for which the number of associated data is greater than i, and:

$$inf = argmin_{c_i \in C}|c_i|, sup = argmax_{c_i \in C}|c_i| \tag{7}$$

2.2 Cluster labeling and content validation

Complementary to overall clustering model evaluation, the role of clusters labeling is to highlight the peculiar characteristics or properties of the clusters associated to a clustering model at a given time. Labeling can be thus used both for visualizing or synthesizing clustering results [13] and for validating or optimizing learning of a clustering method [1]. It can rely on endogenous data properties or on exogenous ones. Endogenous data properties represent the ones being used during the clustering process. Exogenous data properties represent either complementary properties or specific validation properties. Some label relevance indexes can be derivated from our former quality indexes using a probabilistic approach. The *Label Recall L-R* derives directly from Eq. 4. It is expressed as:

$$\underset{c}{L-R}(p) = \overline{W}_c^p \tag{8}$$

The *Label Precision P-R* can be expressed as:

$$\underset{c}{L-P}(p) = \frac{\sum\limits_{d \in c} W_d^p}{\sum\limits_{p' \in d, d \in c} W_d^p} \tag{9}$$

Consequently, the set of labels L_c that can be attributed to a cluster c can be expressed as the set of endogenous or exogenous cluster data properties which maximize the *Label F-measure* that combines the *Label Recall* (Eq. 8) and *Label Precision* (Eq. 9) in the same way than the supervised F-measure described by (Eq.1) would do. As soon as *Label Recall* is equivalent to the conditional probability $P(c|p)$ and *Label Precision* is equivalent to the conditional probability $P(p|c)$, this former labeling strategy can be classified as an expectation maximization approach with respect to the original definition given by Dempster and al. [4].

3 Experimentation and Results

To illustrate the behavior of our new quality indexes, and to compare it to the one of the classical inertia indexes, our test dataset is build up from is a set of bibliographic records resulting from the INIST PASCAL database and covering one year of research performed in the French Lorraine area. The structure of the records makes it possible to distinguish the titles, the summaries, the indexing keywords and the authors as representatives of the contents of the information published in the corresponding article. In our experiment, the research topics associated with the keywords field are solely considered. Our test dataset represents a dataset of 1341 records. A frequency threshold of 3 being finally applied on the index terms, it resulted in a data description set of 889 indexing keywords. These keywords cover themselves a large set of different topics (as far one to another as medicine from structural physics or forest cultivation ...). Moreover, they comprise a high ratio of polysemic forms (like age, stress, structure, ...) that are used in the context of many different topics. The resulting experimental dataset can thus be considered as a complex dataset for clustering.

To carry out the clustering, we exploited in parallel the SOM fixed topology neural method [11], the Neural Gas (NG) free topology neural method [16] and the classical K-means method [15]. For each method, we do many different experiments letting varying the number of clusters from 9 to 324 clusters, employing the size of an increasing square SOM grid as a basic stepping strategy. In the next paragraphs, for the sake of clarity, we dont specifically report the results of K-means because they are similar to those of NG.

The analysis of the results performed by an expert showed that only the SOM method provided homogeneous clustering results on this dataset. Hence, in the case of the NG (or K-means) method, the analyst highlighted the presence of "garbage" clusters attracting most of the data in parallel with "chunk" clusters representing either marginal groups or unformed topics. This behavior, that corresponds to the case of degenerated clustering results due to the dataset clustering complexity, can also be confirmed when one looks to the labels that can be extracted from the clusters in an unsupervised way using the expectation maximization methodology described in section 2.2. Hence, it permits to highlight that the NG mainly produced a "garbage" cluster with very big size that collects more than 80% of the data and attracts (i.e. maximize) many kinds of different labels (730 labels among a total of 889). Conversely, the good results of the SOM method can be confirmed in the same way. Hence, cluster labels extraction also shows that this latter method produces different clusters of similar size attracting semantically homogeneous labels groups, which figure out the main research topics covered by the analyzed dataset.

On the one hand, the results presented in Figure 1A illustrate the fact that the classical indexes of inertia have an unstable behavior which does not make it possible to clearly identify an optimal number of clusters in both contexts of

SOM and NG methods. On the other hand, it also appears in Figure 1B that the behavior of the *Macro- Recall/Precision* indexes is stable and makes it possible to identify an optimal number of clusters in all cases. Indeed, this optimal clusters number can be found out at the break-even point between the *Macro-Recall* and the *Macro-Precision* values (i.e. 100 clusters for NG and 256 clusters for SOM in Figure 1B).

Nonetheless, none of these former groups of indexes, whenever it is solely considered, permits to correctly estimate the quality of the results. Those do not make it possible in particular to discriminate between homogeneous results of clustering (SOM) and degenerated ones (NG or K-means). In both cases, they even present the important defect to privilege this last family of results, illustrating a contradictory behavior (i.e. the worst results are identified as the best, and conversely). In the case of degenerated results, one potential explanation of the better values of aforementioned indexes is that the joint presence of a big amount of "chunk" clusters which are both coherent and necessarily distant of a small amount of "garbage" clusters can compensate, and even hide, the imprecision of these latter because of the cluster-based averaging process performed by those indexes.

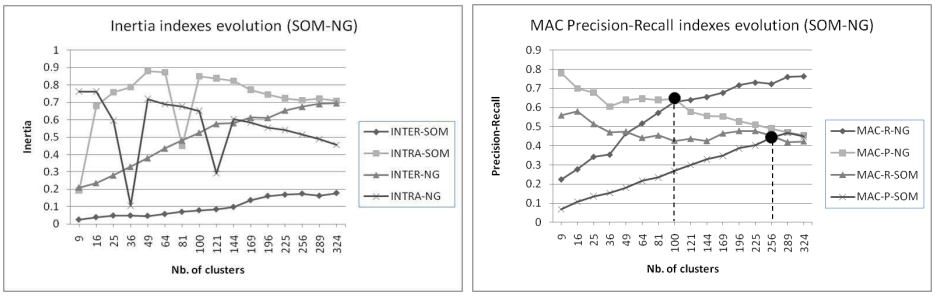

Fig. 1. Inertia (1A) and Macro Recall Precision (1B) indexes evolution as regards to the number of clusters.

In the context of our approach, the detection of degenerated clustering results can although be achieved in two different ways:

The first way is the joint exploitation of the values provided by our Macro- and Micro- Precision indexes, as it is shown in Figure 2A and Figure 2B. The Micro-Recall/Precision indexes have general characteristics similar to the Macro-Recall/Precision. However, by comparing their values with those of the latter indexes, it becomes possible to identify heterogeneous results of clustering. Indeed, in this last case, the Precisions of the clusters of small size will not compensate for any more those of the clusters of big size, and the imprecise properties present in the latter, if they prove to be heterogeneous, will have a considerable effect on

the Micro-Precision. Thus, in the case of NG the differences between the values of Micro- and Macro-Precision are increasingly more important than in the case of SOM, whatever the considered number of clusters (Figure 2A). It proves that the peculiar properties of the clusters in the partitions generated by NG (or K-means) are largely less precise than those of the clusters produced by SOM. The analysis of the evolution of the Micro-Precision curves of the two methods according to the size of the clusters (Figure 2B) permits to clearly highlight that this phenomenon affect more particularly the NG clusters of big size.

A second way to appropriately estimate the quality of clustering results is thus to directly exploit the results provided by the indexes of Cumulated Micro-Precision (CP_m) that focuses on the imprecision of big sized clusters (Eq. 10). In the case of NG, the value of Cumulated Micro-Precision remains very low, regardless to the expected number of clusters (Figure 3A). This is mainly due to the influence of Micro-Precision of "garbage" clusters with significant size that can never be split into smaller groups by the method. In a complementary way, whatever the method considered, the index of Cumulated Micro-Precision ensures accurate monitoring of the quality depending on the chosen configuration in terms of number of clusters. In the case of SOM, the quality loss occurring for some grid sizes (eg. 244 clusters model, corresponding to a 12x12 grid, in Figure 3A) that induces the formation of large heterogeneous clusters is accurately characterized by highly decreasing values of this index. Figure 3B finally illustrates the interest of correcting the Cumulated-Micro-Precision index (Eq. 10) by factorizing it with the ratio of non empty clusters for detecting the optimal number of clusters. The curve associated with this corrected index shows a plateau whose starting point (eg. 256 clusters model in Figure 4B) permits to identify the most efficient partition. In this case, such point highlights that no quality progress can be obtained with higher number of clusters. This information is compliant with the one obtained with Macro-Recall and Macro-Precision indexes (see Figure 1B and associated comments) and thus validates the choice of such point as the characteristic value of the clustering results for a given method.

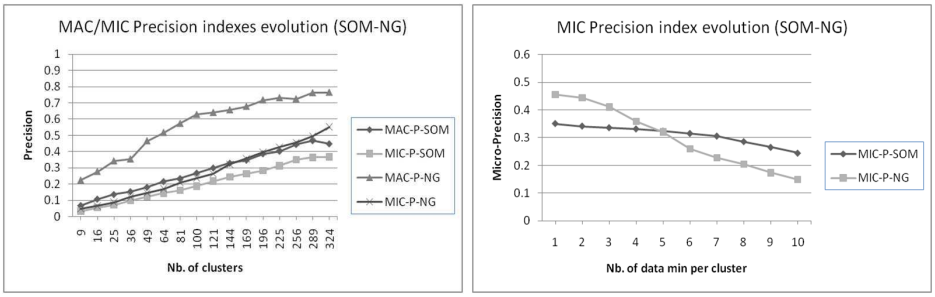

Fig. 2. Evolution of the values of the Micro-Precision (MIC-P) and Macro-Precision (MAC-P) indexes according to the number of clusters (2A) and their size (2B).

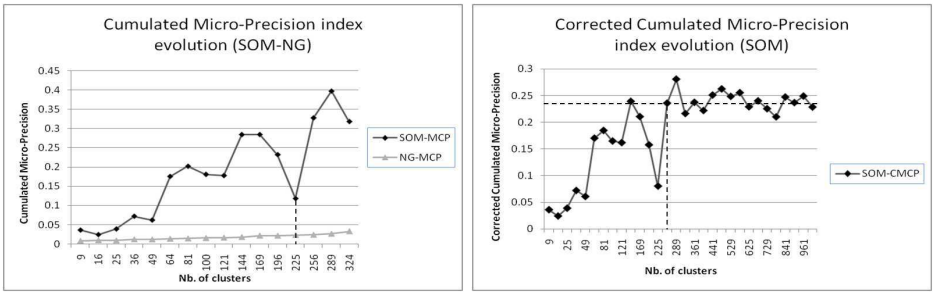

Fig. 3. Evolution of the values of Cumulated Micro-Precision (3A) and Corrected Cumulated Micro-Precision (3B) indexes according to the number of clusters.

4 Conclusion

We have proposed a new approach for the evaluation of the quality of clustering based on the exploitation of the properties associated with the clusters through the indexes of Macro- and Micro- Recall/Precision and their extensions. We have shown the advantage of this approach with respect to traditional of evaluation of clustering quality based on distances, at the same time, by justifying its theoretical basis through its relationship with the symbolic classification approaches, and by showing practical results for the optimization of the number of clusters of a given method. Our experimental have been achieved in a realistic context constituted by a complex textual dataset. In such context, we have shown that our new indexes can accurately assess the global quality of a clustering result while giving the additional possibility to distinguish clearly between homogeneous and degenerated clustering results. We have also shown that our approach can apply to the comparison of the results issued from different methods, as well as to the fine-grained analysis of the results provided by a given method, avoiding in both cases to lead to clustering quality misinterpretation. We have finally shown, through our experiments, the additional capabilities of our approach for synthesizing and labeling the clusters content and we have yet proved their usefulness for a better understanding of the nature of the clustering results. We more specifically tried out our methodology on textual data, but it proves sufficiently general to be naturally applicable on any other type of data, whatever is their nature.

References

1. M. Attik, S. Al Shehabi and J-.C. Lamirel, "Clustering Quality Measures for Data Samples with Multiple Labels", *IASTED International Conference on Artificial on Databases and Applications (DBA), pages 50-57, Innsbruck, Austria, February 2006.*

2. *H.-H. Bock, "Probability model and hypothese testing in partitionning cluster analysis", In: Clustering and Classification, P. Arabie, L.J. Hubert, & G. De Soete (Eds), World Scientific, Singapore (1996), pages 377-453.*
3. *D. Davies, and W. Bouldin, "A cluster separation measure", IEEE Transaction on Pattern Analysis and Machine Intelligence, 1:224-227, 1979.*
4. *A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood for incomplete data via the em algorithm", Journal of the Royal Statistical Society, B-39:1-38, 1977.*
5. *J.Dunn, "Well Separated clusters and optimal fuzzy partitions", Journal of Cybernetics, 4:95-104.*
6. *D. Forest, "Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et danalyse thématique de documents textuels non structurés", PhD Thesis, Quebec University, Montreal, Canada, 2007.*
7. *M. Ghribi, P. Cuxac, J.-C. Lamirel, and A. Lelu, "Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots-clés", Atelier EvalECD2010, Hamamet, Tunisie, January 2010.*
8. *A. D. Gordon, "External validation in cluster analysis", Bulletin of the International Statistical Institute, 51(2), 353-356 (1997), Response to comments. Bulletin of the International Statistical Institute 51(3), (1998), 414-415.*
9. *M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques", Journal of Intelligent Information Systems, 17:2/3, (2001), 147155.*
10. *R. Kassab, and J.-C. Lamirel, "Feature Based Cluster Validation for High Dimensional Data", IASTED International Conference on Artificial Intelligence and Applications (AIA), pages 97-103, Innsbruck, Austria, February 2008.*
11. *T. Kohonen, "Self-organized formation of topologically correct feature maps", Biological Cybernetics, 43:56-59, 1982.*
12. *J-C. Lamirel, S. Al-Shehabi, C. Francois, and M. Hofmann, "New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping", Scientometrics, 60;445-562, 2004.*
13. *J-C. Lamirel, and M. Attik, "Novel labeling strategies for hierarchical representation of multidimensional data analysis results", IASTED International Conference on Artificial Intelligence and Applications (AIA), Innsbruck, Austria, February 2008.*
14. *L. Lebart, A. Morineau, and J.P. Fenelon, "Traitement des données statistiques", Dunod, Paris, 1979.*
15. *J. MacQueen, "Some methods of classification and analysis of multivariate observations", In Proc. 5th Berkeley Symposium in Mathematics, Statistics and Probability, volume 1, pages 281-297. Univ. of California, Berkeley, USA, 1967.*
16. *T. Martinetz and K. Schulten, "A neural gas network learns topologies", Articial Neural Networks, pages 397-402, 1991.*
17. *G.W. Milligan, and M.C. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set", Psychometrika, 50:159-179.*
18. *P.J Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, 20:53-65.*
19. *G. Salton, "The SMART Retrieval System: Experiments in Automatic Document Processing", Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971.*
20. *C. J. Van Rijsbergen, "Information Retrieval", Butterworths, London, England, 1979.*