Involving Validity Indices in Document Clustering

Ahmad El Sayed, Hakim Hacid, Djamel Zighed

Laboratoire ERIC - Université de Lyon
5, avenue Pierre Mendès-France
69676 Bron cedex - France
{asayed, hhacid, dzighed}@eric.univ-lyon2.fr

RÉSUMÉ.

ABSTRACT. *The goal of any clustering algorithm is to find the optimal clustering solution with the optimal number of clusters. In order to evaluate a clustering solution, a number of validity indices are used during or at the end of a clustering process. They can be internal, external or relative. In this paper, we provide two main contributions: First, we present an experimental study comparing the major relative indices in the context of document agglomerative clustering. The objective is to highlight the limits of the existing indices for identifying both the optimal clustering solution and the optimal number of clusters in real datasets. Second, we explore the feasibility of using the relative indices as stopping criteria in agglomerative clustering algorithms. We present a new method that enhances the clustering process with context-awareness to improve their reliability for such utilization.*

MOTS-CLÉS :

KEYWORDS: *Document clustering, agglomerative algorithm, validity indices, context-awareness.*

1. Introduction

Organizing a large set of documents into a number of clusters can highly improve the efficiency and the effectiveness of many text-based applications requiring fast and high-quality navigation and access to their huge documents collection [STE 00]. Over the past decade, a large number of clustering algorithms has been proposed in the literature [BER 02]. These algorithms can be classified along many criteria. For instance, regarding their methodology, they can be either hierarchical, partitionnal, grid-based, or density-based ; Regarding their final output, they can provide either flat or hierarchical clustering ; Regarding the nature of the membership function, they can be either hard (crisp) or soft (fuzzy).

In this paper, our concern relates to the crisp hierarchical agglomerative algorithms providing flat clustering. These algorithms start by initiating each document in a singleton cluster and then repeatedly merge the closest pairs of clusters until all clusters have been merged into a single cluster that contains all documents. The produced result can thus be visualized as a dendogram, that can be cutted at a specific level to keep only the partition of disjoint (flat) clusters.

One of the typical requirements for a good clustering technique in data mining is a minimal input parameter [J. 00]. However, most current clustering algorithms require several key parameters (e.g. number of clusters) and they are thus not practical for use in real world applications. This is a difficult and often an ill-posed problem since users often ignore the optimal parameters, and thus the final results depend on subjectively chosen parameters that do not necessarily fits the dataset. Roughly, the goal of any clustering algorithm is two folds : (1) Finding the optimal clustering solution (i.e. quality of the resulting clustering), and (2) the optimal number of clusters. One option to avoid the need for parameters, is to evaluate the quality of different clustering solutions along different number of clusters, in order to finally choose the solution giving the best results. In cluster analysis, the procedure of evaluating the results is known under the term cluster validity, and the indices that aims at comparing different solutions with different parameters are known as relative validity indices [HAL 02b].

In this paper, we provide two main contributions. Firstly, due to their high computational cost, relative indices are most often studied in a two or three dimensional datasets [HAL 02b, MIL 85, DUN 74, DAV 79]. This surely allows a better vizualisation of results, but do not reflect the reality of their performance in real world applications where data is often multidimensional. Thus, we have chosen to study their performance in a widely used task : document agglomerative clustering. The study includes the existing indices and a new index that we add to the list. Indices are compared according to their ability to identify both the optimal clustering solution and the optimal number of clusters.

Secondly, in literature, the determination of the optimal solution is performed a-posteriori, after evaluating the different solutions obtained with the different number of clusters [RAS 99, HAL 02b, MIL 85, TIB 00]. However, with such an approach, an agglomerative algorithm needs to go until the end of the clustering process (i.e. the

root cluster) before identifying the optimal solution, which is indeed a time and an effort waste. Along those lines, we explore the feasibility of using the relative indices as stopping criteria in agglomerative clustering algorithms. We propose a new method that aims at enhancing the clustering process with context-aware decisions taken along with validity indices. Experimental results show that the method is a step-forward in using, with more reliability, validity indices as stopping criteria.

The remainder of the paper is organized as follows. After a brief overview on cluster validity indices in the next Section, we describe in Section 3 our context-aware method to enable the usage of these indices as stopping criteria. Experimental study on four document collections is presented in Section 4. We conclude in Section 5 by summarizing and drawing some future works.

2. Cluster Validity Indices

In general, we can distinguish two families of cluster validity indices : A first family of indices compares a clustering solution with an a-priori specified structure [HAL 02a]. In this family, indices can be either external or internal. External indices evaluate a clustering solution by comparing it to an a-priori specified structure that reflects the desired result over the dataset (e.g. *F-Score measure, entropy, Jaccard Coefficient, Rand Statistic*). Internal indices evaluate a solution by comparing it to an a-priori specified structure extracted using only quantities and features inherited from the dataset itself (*CPCC, Hubert τ statistic*). A second family of indices compares a clustering solution to another one obtained with the same algorithm but with different parameters or hypothesis [HAL 02b]. This can help finding the optimal parameters that fits the dataset. Here we can find the so called relative indices. Since our concern in this paper is to find the optimal solution across different numbers of clusters k, our focus in the following will be on relative indices.

Relative validity indices turn around two main points : (i) Maximizing the compactness between elements within the same clusters (intra-cluster), and (ii) Maximizing the separation between elements within distinct clusters (inter-cluster). The usage of these indices differs according to the type of algorithm that tends to optimize them [ZHA 05, DUD 01]. For instance, using a hierarchical agglomerative algorithm, a particular validity index VI is calculated at each level for the different merging possibilities. The pair of clusters that is selected to be merged, is the one that leads to a solution that optimizes VI.

According to their behavior, we can distinguish between two kinds of relative indices[1] : (1) Indices scaling with k, making it hard to identify the optimal k (*CH, KL, H1, H2*). This issue is often resolved by inspecting the "knee" on the graph, by the gap statistics approach [TIB 00], or by the stability approach [BEN 02]. (2) Indices not scaling with k, namely not systematically following the trend of k. In this case, the optimal k is more easily chosen as the point on the graph maximizing/minimizing

1. To check corresponding formulas, readers are invited to follow references.

VI. Due to their facility of interpretation, we choose to focus in the rest of this paper on this last kind of indices. The other motivation is related to our goal in the following section, which is to enable the usage of relative indices as stopping criteria, a hard task that will get much harder if the optimal k must be selected using relatively sophisticated techniques. For this kind of indices, we can find those developed for generic clustering purposes (*Dunn indices [DUN 74, BEZ 97], the Davies-Bouldin (DB) index [DAV 79], (RMSSDT, SPR, RS, CD) [SHA 96], (SD, S_Dbw) [HAL 02b]*), and those developed for document clustering purposes [RAS 99] (*C1, C2, C3, C4*).

We added a new validity index $H3$ to this list ; it is inspired from the $H1$, $H2$ indices proposed by Zhao [ZHA 04]. The difference is that $H3$ does not follow the trend of k after having removed its sensitivity to k in an ad-hoc manner. It is computationally less expensive than the other relative indices, since it deals with a collection centroid to calculate the inter-cluster separation (i.e. complexity $O(N)$), while other indices mostly see the inter-cluster as a pairwise similarity between clusters (i.e. complexity $O(N^2)$).

$$H3(k) = \frac{\sum_{i=1}^{k} n_i . \sum_{j=1}^{n_i} sim(D_j, C_i)}{(\sum_{i=1}^{k} sim(C_i, C))/k}$$

where sim denotes similarity between two objects, C_i denotes the centroid of a cluster S_i containing n_i elements, D_j denotes the vector of document d_j, and C denotes the collection centroid which is the average vector of all cluster's centroids.

3. Exploring Indices Usage as Stopping Criteria

3.1. *Problem Definition*

As mentioned earlier, the classical usage of relative validity indices for determining the k yielding the optimal clustering solution comes a-posteriori, *after* evaluating the different solutions provided by a clustering algorithm throught all the possible k values [MIL 85, RAS 99, HAL 02b]. However, in agglomerative algorithms, once reaching the flat optimal solution at $k = \alpha$, all the remaining actions (until $k = 1$) are obviously a time and an effort waste because we will end up by considering the solution provided at α. Hence, finding a relevant stopping criterion is primordial. In litterature, stopping criteria rely, in most cases, on input user parameters. For instance, in agglomerative algorithms, these parameters can be a predefined number of clusters, a minimum similarity between clusters, a maximum similarity gap between successive levels, etc. This kind of stopping criteria have serious limitations since users often ignore the parameters that best fit the dataset[2].

2. The interest here is in hard clustering algorithms. However, many stopping criteria are defined quantifying the degree to which a model fits a dataset in probabilistic clustering algorithms. Readers are invited to see [RIS 89, FRA 98] for examples of such criteria.

A challenging approach to address this issue is to make use of relative indices in order to develop an incremental agglomerative algorithm [DUD 01] able to stop once reaching the "right" optimal solution in terms of a validity index at $k = \alpha$. An intuitive approach is thus to go on with a clustering process while improving a specific index in a stepwise fashion, and to stop once reaching a point $k = \beta$ where no further (significant) improvement[3] can be done with any (merging) action. However, such an ad-hoc approach suffers from ignoring, at the specific level β, whether it has truly reached the optimal solution (i.e. $\alpha = \beta$) or a better solution will come afterward if it accepts a quality decrease at β. The major problem is that validity indices are using too much local information to take a global decision, e.g. stopping the process.

3.2. *Our Method*

3.2.1. *Context-Aware Clustering*

As one could notice, addressing the described issue is a tough and challenging task. Along those lines, we developed a method that aims at enhancing the clustering process with context-aware decisions taken along with validity indices. The end-goal is to enable the usage of validity indices as stopping criteria where a First Drop (FD) in the quality of a clustering solution can more relevantly indicates reaching the optimal solution. The underlying idea is to provide clustering algorithms with a wider vision on the dataset partition, which will enable them to take decisions while having in ”mind” an ”idea” on what could happen next if a specific action is undertaken. As we are seeking the hierarchical agglomerative algorithms, the method applies the following heuristic at each level of the process : consider the M closest pairs of clusters, then estimate the VI after trying to merge each of the M pairs. Among the mergings that improve VI, merge the pair with the lowest Context Risk (CR). If no merging improves VI, merge the pair optimizing VI.

Before merging any two clusters, the method examines the context of the resulting new cluster candidate S_p in terms of its K Nearest Neighbors KNN (i.e. surrounding clusters). Suppose that assessing S_p as a new cluster optimizes VI. If the context tells, though, that creating S_p can lead to a global quality degradation in terms of VI in next iterations, the method chooses another S_q improving VI at a minimal context risk. This surely implies a slower improvement in VI, but has the advantage of continuously pushing, as much as possible, risky merging actions entailing possible future degradations for later processing. Taking the "safest" action at each level leads an expected degradation to occur as late as possible during the process. Thus, a First Drop (FD) in the clustering solution quality is likely to occur closer to the optimal solution, which will offer the possibility to the algorithm to consider more relevantly FD at k as a stopping criteria, and the solution provided at $(k - 1)$ as the optimal clustering solution.

3. While a significant improvement is required with indices scaling with k, a slight improvement is enough for indices not scaling with k.

Note that calculating VI for all the possible mergings between N clusters will lead to a high complexity of $O(N^2)$ at *each level* of the process. We overcome this by considering, at a given level, only the M closest pairs of clusters ($M = 10$ in our experiments), since they form the most potential candidates to improve VI.

3.2.2. *Context Space Composition*

For each new cluster candidate S_p, a CR expresses how risky can be assessing S_p for the overall clustering quality in the expected upcoming mergings given the context of S_p. Consider the two new clusters candidates S_p and S_q illustrated respectively in Figures 1 and 2 with five context clusters each ($S1...S5$). We assume that S_p, with its KNN neither too close nor too distant from its centroid, is more risky than S_q, with its KNN either too close or too distant from its centroid. More precisely, we decompose the context space of a new cluster candidate S_p into three layers using four thresholds $t0$, $t1$, $t2$, $t3$:

1) *Intra layer* Clusters within this layer reduce CR as they should not lead to a quick drop in VI. For this, they have to be close enough to S_p, therefore, likely to be merged with S_p in next iterations without causing a significant degradation (comparing to the r previous mergings) in the global intra-cluster compactness. As a matter of fact, the clusters are getting larger over mergings, and thus the intra-cluster is continuously degradating. At a level k where FD did not occurred yet, we suppose that all the previous mergings that caused degradations in the intra-cluster are acceptable. According to this intuition, this layer is delimited by the thresholds $t0 = 0$ and $t1$ which is defined as the radius of a new cluster candidate S_p augmented by the standard deviation of radius values obtained following the r previous mergings[4]. A radius is the maximum distance between the centroid of S_p and an element within S_p.

$$t1(S_p) = radius(S_p) + StDev(radius(S_{k-r}...S_{k-1}))$$

2) *Inter layer :* Clusters within this layer reduce CR as they should not lead to a quick drop in VI. For this, they have to be distant enough from S_p, therefore, not likely to be merged with S_p in next iterations, and keeping them outside would contribute to improve (or at least not to degradate) the global inter-cluster separation ; This layer is delimited by a first threshold $t2$ that we define as the average pairwise inter-cluster distance between the KNN of S_p, reduced by the standard deviation of its homologous values obtained following the r previous mergings. Getting the average separation between clusters surrounding S_p, will give a hint on the minimum required inter-distance to improve the local inter-cluster separation around S_p, which will most likely improve the global inter-cluster separation. $t2(S_p)$ is calculated as follows :

$$t2(S_p) = AvgInter(S_p) - StDev(AvgInter(S_{k-r}..S_{k-1}))$$

4. We fixed r to 10 in our experiments.

$$AvgInter(S_p) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} dist(S_i, S_j)}{K.(K-1)/2} \quad i \neq j$$

We decided to set the same margin m for the intra and inter layers in order to have a balanced scores in both layers. Subsequently, the other inter-layer threshold $t3$ is defined by $t3 = t2 + t1$, which implies $m = t1 = t3 - t2$

3) *Risk layer :* Clusters within this layer increase CR because we consider that they could lead to a fast drop in the global clustering quality, whether on the inter-cluster or intra-cluster level. Actually, these clusters, if merged with S_p in next iterations, would contribute to a significant degradation in the intra-cluster compactness since they are not enough close to S_p, and if not merged with S_p in next iterations, would not contribute to any significant amelioration in the inter-cluster separation since they are not enough distant from S_p. This layer is delimited by the tresholds $t1$ and $t2$ previously defined.

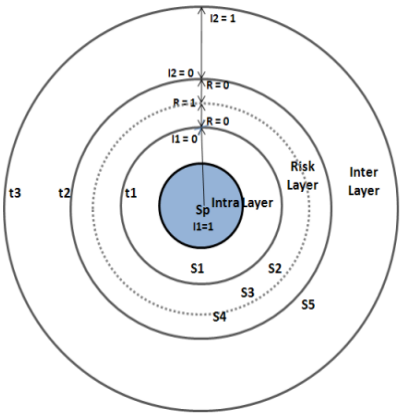

Figure 1. *The three-layers context space of the new risky cluster candidate S_p*

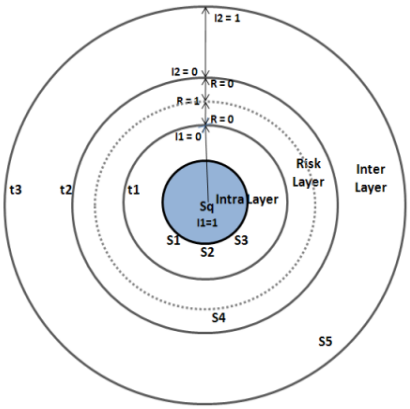

Figure 2. *The three-layers context space of the new non-risky cluster candidate S_q*

3.2.3. *Context Risk Calculation*

Finally, to calculate CR for S_p, we use the following formula :

$$CR(S_p) = \frac{1}{K}(\sum_{i=1}^{n1} R(S_i, S_p) - \sum_{j=1}^{n2} I1(S_j, S_p) - \sum_{h=1}^{n3} I2(S_h, S_p))$$

$R(S_i, S_p)$, $I1(S_j, S_p)$, $I2(S_h, S_p)$ denote the score given for a cluster S situated respectively in the risk layer, intra layer, and inter layer. K refers to the predefined number of nearest neighbors, which we fix to 10 in our experiments. $n1$, $n2$, $n3$ denote

the number of clusters situated respectively in the risk, intra and inter layers. All the scores are distributed along a [0,1] range according to their distances with the centroid of S_p (See Figures 1 and 2). Consequently, CR varies between -1 (for a minimal risk) and 1 (for a maximal risk). For a contextual cluster S_x, having a distance d_x with the centroid of S_p, its score is calculated with respect to the following conditions :

$$\left\{ \begin{array}{l} if\ d_x < t_1 \Rightarrow I1(S_x, S_p) = \frac{t_1 - d_x}{m} \\ else\ if\ d_x > t_2 \Rightarrow I2(S_x, S_p) = \frac{d_x - t_2}{m} \\ else\ if\ d_x > t_3 \Rightarrow I2(S_x, S_p) = 1 \\ else\ if\ t_1 \leq d_x < (t1 + t2)/2 \Rightarrow R(S_x, S_p) = \frac{d_x - t_1}{m/2} \\ else\ if\ (t1 + t2)/2 \leq d_x \leq t2 \Rightarrow R(S_x, S_p) = \frac{t_2 - d_x}{m/2} \end{array} \right\}$$

4. Experimental Study

4.1. *Document Collections*

For our experimental study, we used a total of four datasets, whose general characteristics are summarized in Table 1. These are four distinct datasets extracted from the Reuters corpus[5] containing a set of documents being assigned by an expert to a single topic each (e.g. *economy, politics, sports*). We avoided multi-topics documents since we are dealing with crisp algorithms where a document can belong to only one cluster. The study focuses on a set of indices not scaling with k for the reasons mentioned in Section 2 (*DB [DAV 79], Dunn [DUN 74], m-Dunn [BEZ 97], (C1, C2, C3, C4) [RAS 99], H3*). Due to time constraints, we limited our experiments on DS3 and DS4 to the *H3* index since it is computationally more affordable on large datasets than the other indices.

The vector-space model [SAL 89] is used to represent a document d by a vector v in a multidimensional space, where each dimension represents a term expressed by its $tf.idf$ weight. A cluster is represented by its centroid C, which is the average of documents vectors $\bar{V}$ contained in the cluster. To capture similarity between two clusters, we use the cosines similarity measure between the two cluster's centroids after normalizing each centroid C to be of unit length ($||C_{tfidf}||$ = 1). The similarity formula is then : $sim(C_i, C_j) = cos(C_i, C_j) = C_i^t C_j$. To calculate distances, we use : $dist(C_i, C_j) = 1 - sim(C_i, C_j)$.

4.2. *Evaluating Relative Validity Indices*

Having already classified document collections, one efficient way for evaluating relative indices is to compare their behaviors with those of external indices which we suppose bear the optimal behaviours since they are based on a predefined structures set by experts. Therefore, we run the agglomerative algorithm with a given index on

5. Reuters corpus, volume 1, English language, release date : 2000-11-03

Dataset	# of documents	# of topics
DS1	100	22
DS2	200	24
DS3	500	38
DS4	700	53

Tableau 1. *Summary of datasets used for our experiments.*

a given dataset. Each solution provided at each level k of the process is evaluated by means of the target relative index (predicted quality) and the *F-Score measure* (real quality). As in [ZHA 05], the *F-Score* is calculated by first identifying for each class of documents the cluster that best represent it, and then measuring the overall quality of a solution by the average of the different classes qualities. Following this procedure, we study the ability of each relative index to reach the predefined structure, in terms of identifying both the optimal clustering solution and the optimal number of clusters. Note that these two goals do not necessarily overlap. Actually, since algorithms are error-prone, a "real" optimal solution can lie under a number of clusters different from the "real" optimal one.

4.2.1. *On identifying the Optimal Clustering Solution*

Firstly, we test the relative indices ability for evaluating a given clustering solution in order to identify the optimal one. Approving the $F\text{-}Score$ output as the "Gold Standard" output at each iteration, we present in Figure 3 the indices results evaluated from three different angles :

– Their correlation with the $F\text{-}Score$: By studying correlation between predicted values and real values, we can figure out to which extend a relative index can behave similarly to an external index.

– The optimal $F\text{-}Score$ reached across the different k : It represents the optimal clustering quality that a VI can reach if it shares the same optimal k with $F\text{-}Score$. Values express also to which extend, (merging) actions based on a given index can lead to correct/incorrect partitions among clusters.

– The $F\text{-}Score$ reached at the optimal value of VI : This is a good indicator of the overall solution quality that a VI can reach. By comparing these values to the previous values (i.e. optimal *F-Score*), one can check to which extend the optimal solution provided by a VI is close to the real optimal solution.

4.2.2. *On Identifying the Optimal Number of Clusters*

Secondly, we test the relative indices ability for identifying the optimal k which we define as the number of distinct topics in a dataset. We present in Figures 4 the indices results evaluated from two different angles :

– k at the optimal value of VI, which represents to which extend a VI, with its actual trend for determining the optimal k, is able to approach the real optimal k value.

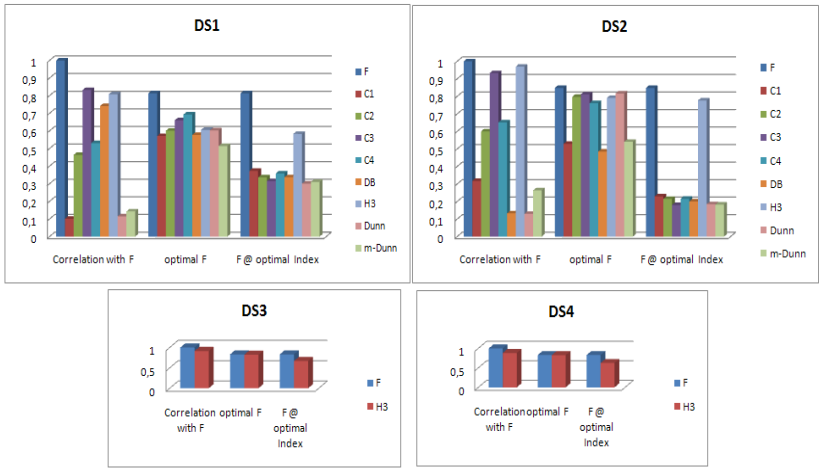

Figure 3. *Indices ability to identify the optimal clustering solution in each dataset*

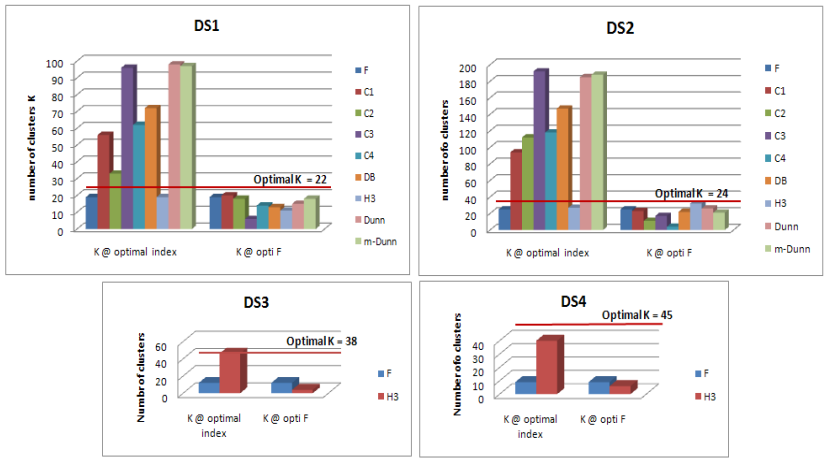

Figure 4. *Indices ability to identify the optimal number of clusters in each dataset*

– k at the optimal value of the *F-Score*, which represents to which extend a VI, if it had the trend of $F\text{-}Score$ for determining the optimal k, is able to approach the real optimal k value.

4.2.3. *Discussion*

After examining the different graphs in Figures 3 and 4, we can conclude the following remarks :

First, we can say that most relative indices provide a satisfying performance for evaluating solutions in comparison to our external index. Actually, they behave enough closely to the external index (Correlation with F in Figure 3), and are also able to attain a maximal *F-Score* comparable to the optimal one attained by the $F\text{-}Score$ measure itself runned separately (Optimal F in Figure 3). In addition, as one could expect, indices developed for generic purposes (e.g. *DB, Dunn, m-Dunn*) are among the indices giving the worth results for evaluating solutions. This can be explained by their weak ability to deal with the sparse nature of our datasets which usually require more representative quality estimation for the global partition.

Second, indices show some "rigid" trends for detecting the optimal k over different datasets. If we look at the optimal k in terms of each index (K @ optimal index) in Figure 4, we can notice that *C3, C4, H3, DB, Dunn, m-Dunn*, although not scaling systematically with k, are keeping similar relative trends towards the optimal k over DS1 and DS2. This is indeed an important shortcoming since the optimal k is supposed to depend uniquely on the dataset. Nevertheless, if they had flexible trends like the *F-Score*, they would show much more variations in defining their optimal k (K @ opti F), showing more significantly how much an index, regardless of its current trend, is able to detect the optimal k. From this point of view, we can conclude that *C1, H3, Dunn* are among the best indices for detecting the optimal k.

Third, basing on the previous remark, we can deduce that the gap between optimal k in terms of each index and the "real" optimal k (K @ optimal Index in Figure 4) is not highly informative vis-a-vis the error rate, since it highly depends on both the current trend of each index and the predefined optimal k that reflects only a certain level of granularity that could be too specific or too generic. From this point of view, the underlying gap can simply inform us on which indices are more suitable for cases requiring rather high k or low k.

Finaly, in spite of their "good" performance for evaluating a solution, most indices are still long way from reaching the "real" optimal solutions. This can demonstrated in Figure 3 by the large gap between the optimal *F-Score* (Optimal F) and the *F-Score* reached at the optimal k in terms of an index (F @ Optimal Index). The gap is related, among others, to the indices difficulty to detect the optimal k. For instance, an index showing a trend to high number of clusters, will provide a bad clustering quality if the optimal solution lies under a small number of clusters. This is the case of $C3$, that gives an optimal solution at $k = 192$ with $F = 0,18$, while the real optimal solution was lying under $k = 17$ and $F = 0.82$. An exception to this gap is noticed with the $H3$ index, whose high ability for reaching the optimal k is surely affecting its high ability for reaching the optimal clustering solutions.

4.3. *Evaluating the Context-Aware Method*

In this section, we explore the added-value of enhancing a clustering process with context-awareness in order to enable validity indices usage as stopping criteria. As

stressed earlier in this paper, the goal is to approach, as much as possible, the solution provided before FD to the optimal solution. We excluded from the following experiments some indices that showed to be inappropriate for the context-aware method because they provide either too unstable curves to be stabilized (e.g. $Dunn$, m-$Dunn$) , or too stable curves in our datasets to show clearly the effect of the context enhancement (e.g. $C3$).

4.3.1. *Approaching the Optimal Number of Clusters*

Firstly, we study to which extend the method allows FD to approach the optimal number of clusters. Therefore, we demonstrate in Figure 5 the complete agglomerative clustering process ($k = n \rightarrow 1$) divided into three parts :

– P1 : This part goes from the initial set ($k = n$) to the last point before FD. Thus, using a VI as a stopping criterion will lead the process to the last point of P1.

– P2 : This part goes from FD to the optimal clustering solution. It represents the part that must be processed but would not if VI is used as a stopping criterion.

– P3 : This part goes from the optimal solution until the root cluster ($k = 1$), which form the unnecessary part that would be performed in vain if a VI is not used as a stopping criterion.

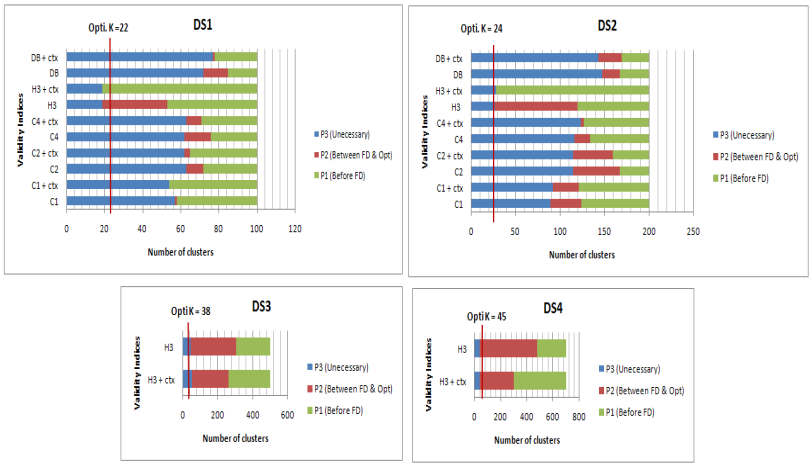

Figure 5. *The added-value of the context-aware method in approaching the optimal k in each dataset*

By observing Figure 5, we can quickly notice the added-value of the context-aware method. On the first hand, it avoids a clustering algorithm from processing all the P3 parts which is a time and an effort waste. On the other hand, it contributes to reduce P2, since in most cases, FD occurs remarkably closer to the optimal solution. This will surely enable us to consider more relevantly a solution before FD as the optimal solution. The contribution is made clear especially when observing the $H3$

index in DS1 : while the optimal solution is found at $k = 19$ with $F = 0.58$, using the index alone entails a first drop to occur at $k = 53$ with $F = 0.33$. However, when adding context-awareness to the process, a first drop occurred exactly at $k = 19$ with $F = 0.60$, which is simply the ideal intended result.

4.3.2. *Quality of the Optimal Solutions*

Since a "context-aware" algorithm is no more taking the merging decisions that optimizes VI, one may imagine that the method, although approaching the optimal k, can deteriorate the quality of the solutions given by an index. However, results in Figure 6 show the opposite ; actually, the graphs illustrating the *F-score* at the optimal value of each VI assess that with the context-aware method, we can still have a comparable and sometimes better clustering quality than the standard method without involving any context-awareness. In average, using the method led the $F\text{-}Score$ at the optimal value of VI to a decrease of 0.06%, 0.10% in DS2, DS3, and to an increase of 3.11%, 1.88% in DS1, DS4 respectively.

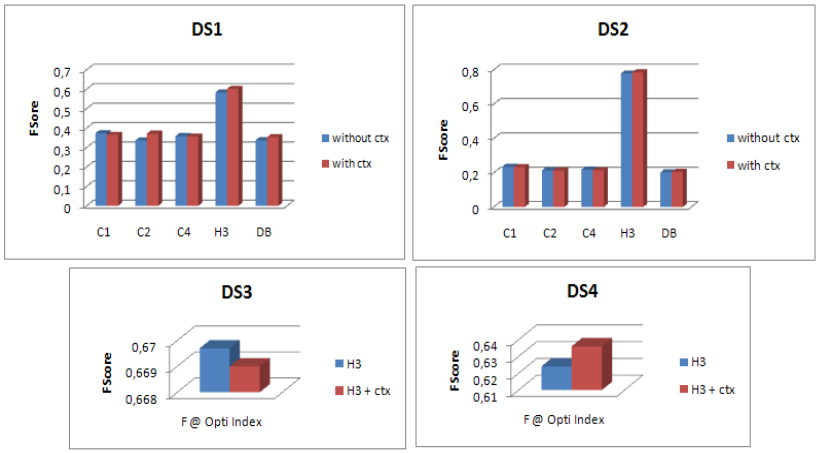

Figure 6. *F-Scores obtained at the optimal solution in terms of VI used with and without context in each dataset*

4.3.3. *Quality of the Final Solutions*

More informative than the quality of the optimal solutions, is the quality of the final provided solutions for the user when stopping before FD. In Figure 7, these solutions, provided with/without using context-awareness, are evaluated also in terms of the *F-Score* measure. In average, using the context-aware method contributed to an $F\text{-}Score$ increase of 14.96%, 20.71%, 8.88%, 21,21% in the DS1, DS2, DS3, DS4 respectively.

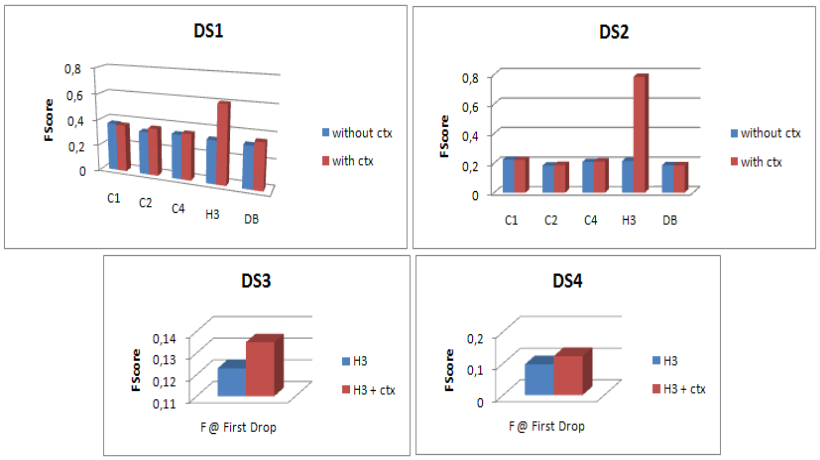

Figure 7. *F-Scores obtained just before FD with and without context in each dataset*

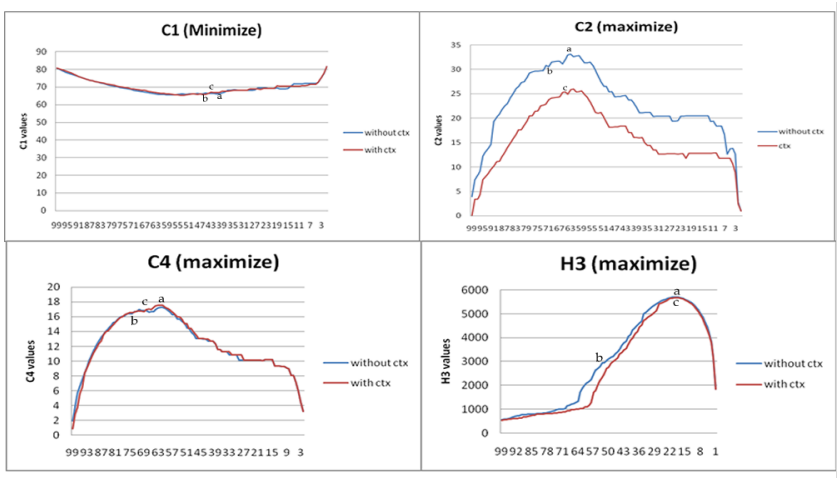

Figure 8. *The indices values along the different number of clusters, with and without using context in DS1*

4.3.4. *Discussion*

In Figure 8, we can see more precisely the effect of enhancing the process with context-awareness in DS1[6], which results in a slower but more reliable evolution in the indices values. For each graph representing an index, we set three points : (a) re-

6. Demonstrations has been limited to DS1 for space constraints.

presents the optimal intended solution, (b) represents the point where an algorithm would stop if no context were involved, (c) represents the point where an algorithm will stop when involving context. By examining the different graphs, we can notice that (c) always approaches considerably (a) comparing to (b), which prove the efficiency of the proposed method. Interestingly, the context-aware method appears to be sensitive to the risk degree of an index. This means that the "level of precaution" that it takes depends on how risky an index is. Thus, when adding context-awareness to the relatively "safe" indices like $C1$ and $C4$, the decisions taken are very close to those taken without context, which results in two similar curves. However, when adding context-awareness for "risky" indices like $C2$ and $H3$, their values evolve much slower than without context, keeping wider gaps between curves.

5. Conclusion and Future Works

On the first hand, we studied the usage relative validity indices in the context of agglomerative document clustering, testing their ability to identify both, the optimal clustering solution and the optimal number of clusters. Experiments performed on four document collections show that most relative validity indices behave enough closely to external indices for evaluating clustering solutions. However, they show some difficulties to detect the optimal number of clusters due to their ”rigid” trends over different datasets. On the other hand, we explored the feasibility of using relative indices as stopping criteria, which is a crucial part for avoiding an algorithm from trying all the possible parameters before assessing the ”right” solution, which is a time an effort waste. We described a new method enhancing an agglomerative algorithm with context-awareness to allow a more reliable usage of validity indices for this purpose. Experimental results show that our method allows an algorithm to considerably approach the optimal solution which is classically identified a-posteriori. As for future works, we aim at evaluating our method with more clustering algorithms and on larger datasets. Given the fact that documents often belong to multiple topics, an application of the method is planned under incremental soft clustering algorithms allowing overlaps between clusters. Furthermore, we believe that a more flexible and significant definition of the context space is still needed to improve results.

6. Bibliographie

[BEN 02] BEN-HUR A., ELISSEEFF A., GUYON I., « A Stability Based Method for Discovering Structure in Clustered Data », *Pacific Symp. on Biocomp.*, 2002, p. 6-17.

[BER 02] BERKHIN P., « Survey Of Clustering Data Mining Techniques », rapport, 2002, Accrue Software, San Jose, CA.

[BEZ 97] BEZDEK J. C., LI W., ATTIKIOUZEL Y., WINDHAM M. P., « A geometric approach to cluster validity for normal mixtures », *Soft Comput.*, vol. 1, n° 4, 1997, p. 166-179.

[DAV 79] DAVIES DL B. D., « A cluster separation measure », *IEEE Trans. on Pat. Anal. and Mach. Int.*, vol. 1(2), 1979.

[DUD 01] DUDA R. O., HART P. E., STORK D. G., « Pattern Classification », *John Willey & Sons*, 2001.

[DUN 74] DUNN, « Well separated clusters and optimal fuzzy paritions », *Journal Cybern*, vol. 4, 1974, p. 95-104.

[FRA 98] FRALEY C., RAFTERY A. E., « How Many Clusters ? Which Clustering Method ? Answers Via Model-Based Cluster Analysis », *Comput. J.*, vol. 41, n° 8, 1998, p. 578-588.

[HAL 02a] HALKIDI M., BATISTAKIS Y., VAZIRGIANNIS M., « Cluster Validity Methods : Part I », *SIGMOD Record*, vol. 31, n° 2, 2002, p. 40-45.

[HAL 02b] HALKIDI M., BATISTAKIS Y., VAZIRGIANNIS M., « Clustering Validity Checking Methods : Part II », *SIGMOD Record*, vol. 31, n° 3, 2002, p. 19-27.

[J. 00] J. H., M. K., « Data Mining : Concepts and Techniques », *Morgan Kaufmann Publishers*, 2000.

[MIL 85] MILLIGAN G. W., COOPER M. C., « An examination of procedures for determining the number of clusters in a data set », *Psychometrika*, vol. V50, n° 2, 1985, p. 159–179.

[RAS 99] RASKUTTI B., LECKIE C., « An Evaluation of Criteria for Measuring the Quality of Clusters », *IJCAI*, 1999, p. 905-910.

[RIS 89] RISSANEN J., « Stochastic Complexity in Statistical Inquiry », *World Scientific Publishing Co.*, 1989.

[SAL 89] SALTON G., *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, 1989.

[SHA 96] SHARMA S., « Applied multivariate techniques », *John Willy and Sonw*, 1996.

[STE 00] STEINBACH M., KARYPIS G., KUMAR V., « A comparison of document clustering techniques », *KDD Workshop on Text Mining*, 2000.

[TIB 00] TIBSHIRANI R., WALTHER G., HASTIE T., « Estimating the number of clusters in a dataset via the gap statistic », 2000.

[ZHA 04] ZHAO Y., KARYPIS G., « Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering », *Machine Learning*, vol. 55, n° 3, 2004, p. 311-331.

[ZHA 05] ZHAO Y., KARYPIS G., FAYYAD U. M., « Hierarchical Clustering Algorithms for Document Datasets », *Data Min. Knowl. Discov.*, vol. 10, n° 2, 2005, p. 141-168.