Evaluation of Internal Validity Measures in Short-Text Corpora*

Diego Ingaramo[1], David Pinto[2,3], Paolo Rosso[2], and Marcelo Errecalde[1]

[1] Development and Research Laboratory in Computacional Intelligence (LIDIC),
UNSL, Argentina
[2] Natural Language Engineering Lab.,
Department of Information Systems and Computation,
Polytechnic University of Valencia, Spain
[3] Faculty of Computer Science (FCC),
BUAP, Mexico
{daingara,merreca}@unsl.edu.ar,
{prosso,dpinto}@dsic.upv.es

Abstract. Short texts clustering is one of the most difficult tasks in natural language processing due to the low frequencies of the document terms. We are interested in analysing these kind of corpora in order to develop novel techniques that may be used to improve results obtained by classical clustering algorithms. In this paper we are presenting an evaluation of different internal clustering validity measures in order to determine the possible correlation between these measures and that of the F-Measure, a well-known external clustering measure used to calculate the performance of clustering algorithms. We have used several short-text corpora in the experiments carried out. The obtained correlation with a particular set of internal validity measures let us to conclude that some of them may be used to improve the performance of text clustering algorithms.

1 Introduction

Document clustering consists in the assignment of documents to unknown categories. This task is more difficult than supervised text categorization [13,8] because the information about categories and correctly categorized documents is not provided in advance. An important consequence of this lack of information is that clustering results cannot be evaluated with typical external measures like F-Measure and, therefore, the quality of the resulting groups is evaluated with respect to *structural properties* or *internal measures*. Classical internal measures used as cluster validity measures include the *Dunn* and *Davies-Bouldin* indexes, new graph-based measures like *Density Expected Measure* and Λ-Measure as well as some measures based on the corpus vocabulary overlapping.

* This work has been partially supported by the MCyT TIN2006-15265-C06-04 project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, pp. 555–567, 2008.

When clustering techniques are applied to collections containing *very short* documents, additional difficulties are introduced due to the low frequencies of the document terms. Research work on “short-text clustering” is relevant, particularly if we consider the current/future mode for people to use ’small-language’, e.g. blogs, text-messaging, snippets, etc. Potential applications in different areas of natural language processing may include re-ranking of snippets in information retrieval, and automatic clustering of scientific texts available on the Web [10].

In order to obtain a better understanding of the complexity in clustering short-text corpora, a deeper analysis of the main factors that have a direct impact on the obtained results is required. Specifically, we are interested in studying whether the internal clustering validity measures are good estimators of the usability of the results from an user viewpoint. For this purpose, several short-text corpora are considered. Since the information about the correct categories of the documents is available, then the quality of clustering results evaluated according to the internal measures can be compared with external ones, in our case, with F-Measure.

Our study is closely related to the work presented in [15] where different internal cluster validity measures are used to predict the quality of clustering results in experiments with samples of the RCV1 Reuters collection [12]. The predicted quality in this case is compared with the real quality expressed by the F-measure values obtained from the classification of a human editor. In our case, we study *very short-text corpora*. The aim is to determine the correlation degree between internal and external clustering validity measures.

The rest of paper is organized as follows. In Section 2 we explain the metrics that are used in the experimental work to determine the quality of the obtained clusters. Section 3 describes the short-text corpora used in the experiments. The experimental results are shown in Section 4. Finally, we draw some conclusions and we discuss the future work.

2 Validity Measures

Cluster validity is a measure of goodness for the results obtained by clustering algorithms. There exist two types of cluster validy measures, namely, external and internal. The difference relies, respectively, on the use or not of a pre-specified structure of the data which is imposed usually by an expert. Following we describe a particular set of internal measures, some of them previously investigated in [15]. We introduce also two new measures based on the vocabulary overlapping whose basics were already presented in [10]. We start below with a short description of the well-known external validity F-Measure we used to calculate the correlation of the obtained results. Other internal validity measures (such as the Silhouette coefficient, correlation, cophenetic distance, purity, Neill’s conditional entropy and Newman’s Q-Measure) could have been explored. For instance, relative closeness and relative interconnectivity were introduced in [5] in the framework of dynamic modeling for hierarchical clustering. However, we consider that the analysis of all of them would be out of the scope of this paper.

2.1 *F*-Measure

In the context of clustering, F-Measure is an external validity measure that combines both, *precision* and *recall*. It may be formally defined as follows. Let D represent the set of documents, $\mathcal{C} = \{C_1, ..., C_k\}$ be a clustering of D and $\mathcal{C}^* = \{C_1^*, \ldots, C_l^*\}$ designate the human reference classification of D. The *recall* of a cluster j with respect to a class i, $rec(i,j)$ is defined as $|C_j \cap C_i^*|/|C_i^*|$. The *precision* of a cluster j with respect to a class i, $prec(i,j)$ is defined as $|C_j \cap C_i^*|/|C_j|$. Thus, the F-measure of the cluster j with respect to a class i is $F_{i,j} = \frac{2 \cdot prec(i,j) \cdot rec(i,j)}{prec(i,j)+rec(i,j)}$ and the overall F-measure is defined as:

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

2.2 The *Λ*-Measure

Let us consider a data collection as a weighted graph $G = \langle V, E, w \rangle$ with node set V (representing documents), edge set E (representing similarity between documents) and weight function $w : E \to [0,1]$ (representing a similarity function between documents). Λ-Measure computes the *weighted partial connectivity* of $G = \langle V, E, w \rangle$. Formally [15], let $C = \{C_1, ...C_k\}$ be a clustering of the nodes V of a weighted graph $G = \langle V, E, w \rangle$, then the Λ internal measure of C is:

$$\Lambda(\mathcal{C}) = \sum_{i=1}^{k} \lambda_i \cdot |C_i| \tag{2}$$

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. It is expected that the higher is the value of Λ the better is the clustering obtained.

2.3 The Density Expected Measure

A graph G $= \langle V, E, w \rangle$ may be called sparse if $|E| = \mathcal{O}(|V|)$, whereas it is called dense if $|E| = \mathcal{O}(|V|^2)$. Then we can compute the density θ of a graph from the equation $|E| = |V|^\theta$ where $w(G) = |V| + \sum_{e \in E} w(e)$, in the following manner:

$$w(G) = |V|^\theta \Leftrightarrow \theta = \frac{\ln(w(G))}{\ln(|V|)} \tag{3}$$

θ can be used to compare the density of each induced subgraph $G^{'} = \langle V^{'}, E^{'}, w^{'} \rangle$ with respect to the density of the initial graph G. G' is sparse (dense) compared to G if $\frac{w(G^{'})}{|V'|^\theta}$ is smaller (bigger) than 1. Formally [15], let $\mathcal{C} = \{C_1, .., C_k\}$ be a clustering of a weighted graph $G = \langle V, E, w \rangle$ and $G_i = \langle V_i, E_i, w_i \rangle$ be the induced subgraph of G with respect to cluster C_i. Then the *Density Expected*

Measure (DEM) $\overline{\rho}$ of a clustering $\mathcal{C}$ is obtained as shown in Eq. (4). A high value of $\overline{\rho}$ should indicate a good clustering.

$$\overline{\rho}(\mathcal{C}) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^{\theta}} \tag{4}$$

2.4 The Dunn Index Family

The Dunn Index Family identifies cluster sets that are compact and well separated. Let $C = C_1, ...C_k$ be a clustering of a set of objects D, $\delta : C \times C \rightarrow \mathbf{R}$ be a cluster to cluster distance and $\varDelta : C \rightarrow \mathbf{R}$ be a cluster diameter measure. Then all measures of the following form are called Dunn indices.

$$I(C) = \frac{min_{i \neq j} \delta(C_i, C_j)}{max_{1 \leq l \leq k} \varDelta(C_i)} \tag{5}$$

For our analysis we have used $\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y)$ and $\varDelta(C_i) = 2 \left(\frac{\sum_{x \in C_i} d(x, c_i)}{|C_i|} \right)$ (see the Bezdek definition [3]), where $d : D \times D \rightarrow R$ is a function that measures distance between objects and c_i denotes the centroid of a cluster C_i. Large values of $I(C)$ correspond to a good cluster partition.

2.5 The Davies-Bouldin Index

This measure combines within-cluster scatter and between-cluster separation of a clustering C. It is obtained as follows.

$$DB(C) = \frac{1}{k} . \sum_{i=1}^{k} R_i(C), \text{ with} \tag{6}$$

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

where $s : C \rightarrow \mathbf{R}$ measures the scatter within a cluster, and $\delta : C \times C \rightarrow \mathbf{R}$ is a cluster to cluster distance (intercluster). For our analysis we defined $s(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} ||x - c_i||$ and $\delta = ||c_i - c_j||$. Small values of DB would correspond to good clusters, since the clusters should be compact and their centers will be far away from each other.

2.6 The Relative Hardness Measure

This measure, introduced first in [10], calculates the vocabulary overlapping degree of a given set of clusters. Although this measure may be used both, as a external or a internal clustering validity measure, here we perform the evaluation of the data from an internal viewpoint. Formally, given a corpus C made up of n categories (CAT), the Relative Hardness (RH) of $C = \{CAT_1, CAT_2, ..., CAT_n\}$ is:

$$RH(C) = \frac{1}{n(n-1)/2} \times \sum_{i,j=1;i<j}^{n} Similarity(CAT_i, CAT_j), \tag{7}$$

where the similarity among categories is obtained by using both, the Jaccard coefficient and the cosine measure in order to determine their overlapping (see Equation (8) and (11), respectively).

$$Similarity(CAT_i, CAT_j) = \frac{|CAT_i \bigcap CAT_j|}{|CAT_i \bigcup CAT_j|} \tag{8}$$

In the above formula we have considered each category i as the "document" obtained by concatenating all the documents belonging to the category i. In Equation (9), w_{ij} is the weight of the term t_j in the category i (CAT_i). idf_j (Eq. (10)) is the inverse category frequency of the term t_j and, finally, the similarity (Eq. (11)) is the cosine of the angle between the categories vectorial representation of a given corpus. We have named the RH calculated with Eq. (8) as RH-J, whereas the one that uses Eq. (11) as RH-C.

$$w_{ij} = tf_{ij} \times idf_j \tag{9}$$

$$idf_j = log\left(\frac{n}{df_j}\right) \tag{10}$$

$$Similarity(CAT_i, CAT_j) = \frac{\sum_k w_{ik} \times w_{jk}}{\sqrt{\sum_k w_{ik}^2} \times \sqrt{\sum_k w_{jk}^2}} \tag{11}$$

3 Data Sets

The aim of this research work was to analyse the behaviour of different clustering internal measures over different short-text corpora of unrelated domains. The datasets used in the experiments carried out are described as follows.

3.1 The CICLing-2002 Corpus

This dataset is made up of 48 abstracts from the *Computational Linguistics* domain, which corresponds to articles presented at the *CICLing 2002* conference. Despite the small size, this collection has been used in differents experiments (see [7,11,2,9]). The distribution and the features of this corpus is shown in Table 1.

Table 1. Main characteristics of the *CICLing-2002* corpus

Category	# of abstracts
Linguistics	11
Ambiguity	15
Lexicon	11
Text Processing	11

Feature	Value
Size of the corpus (bytes)	23,971
Number of categories	4
Number of abstracts	48
Total number of terms	3,382
Vocabulary size (terms)	953
Term average per abstract	70.45

3.2 The R8 Dataset

Reuters-21578[1] has been extensively used in text categorization. In the experiments we have carried out, we have used the R8 subcollection of the Reuters-21578 since it is a single-categorized dataset. The characteristics of this corpus are given in Table 2.

Table 2. Main characteristics of the R8 corpus

Category	Test Docs	Train Docs
trade	102	319
grain	34	78
monex-fx	130	366
crude	140	314
interest	87	202
acq	707	1608
ship	43	121
earn	1076	2831

Feature	Test	Train
Size of the corpus (KBytes)	≈912	≈2,567
Number of categories	8	8
Number of documents	2,319	5,839
Total number of terms	150,430	416,431
Vocabulary size (terms)	9,315	15,648
Term average per document	64.87	71.32

Table 3. Sample of the 100 ambiguous words of the *WSI-SemEval* corpora with their corresponding number of instances

Word	Instances
share	3061
rate	1154
president	1056
people	869
state	689
point	619
part	552
system	520
bill	506
future	496
⋮	⋮

(a) nouns

Word	Instances
say	2702
ask	406
turn	402
feel	398
keep	340
go	305
work	273
do	268
believe	257
start	252
⋮	⋮

(b) verbs

[1] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

3.3 The WSI SemEval Corpora

This corpora was provided by the organizers of the *Evaluating Word Sense Induction (WSI) and Discrimination Systems* task of the SemEval 2007 workshop [1]. The dataset consists of 100 ambiguous words (65 verbs and 35 nouns) taken from the English lexical sample task of the same workshop. The corpora is then composed of 100 data collections, each one, corresponding to a specific ambiguous word. The name of a sample of the ambiguous words (10%) along with the number of their instances are presented in Table 3. A set of average values of the characteristics of this corpus is given in Table 4.

Table 4. Other features of the *WSI-SemEval* corpora

Feature	**Value**
Size of the corpus (bytes)	10,644,648
Number of ambiguous words	100
Number of sentences	27,132
Total number of terms	1,555,960
Vocabulary size (terms)	27,656
Average number of sentences (instances)	271.32
Average vocabulary size	47,65
Term average per sentence	57.34

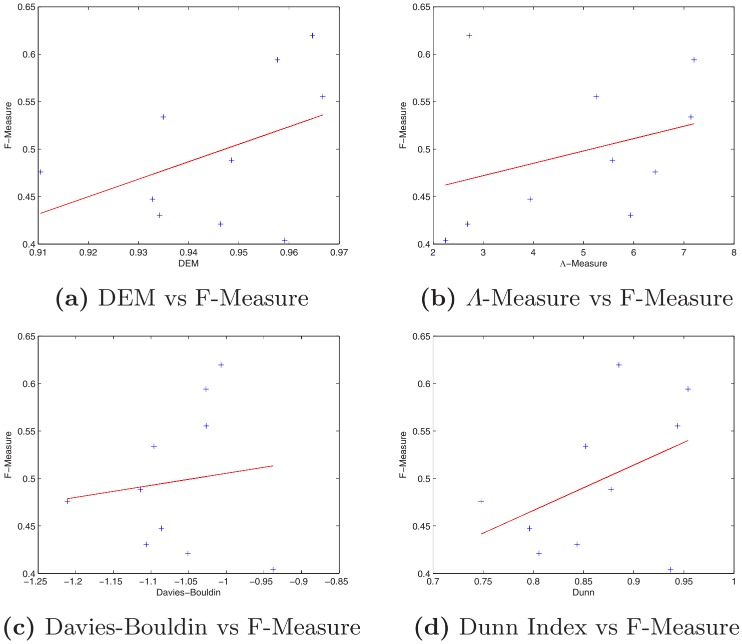
3.4 Subcorpora Generation

We have generated subsets for the *CICLing-2002* and the R8 corpora to analyse the behaviour of the Internal Clustering Validity Measures (ICVM) over all the differents variations of these corpora. We considered all the possible combinations of more than two categories from the corpus and for each of them we calculated its ICVM value. Therefore, for a corpus of n categories, a number of $2^n - (n+1)$ possible subcorpora is obtained.

4 Experimental Results

The aim of this research work was to investigate the possible correlation between the external measure F and several internal clustering measures. We executed the K-Star agglomerative clustering method [14] over the corpora previously mentioned. The F-Measure and all the internal clustering validity measures were evaluated with the clusters obtained by this clustering method.

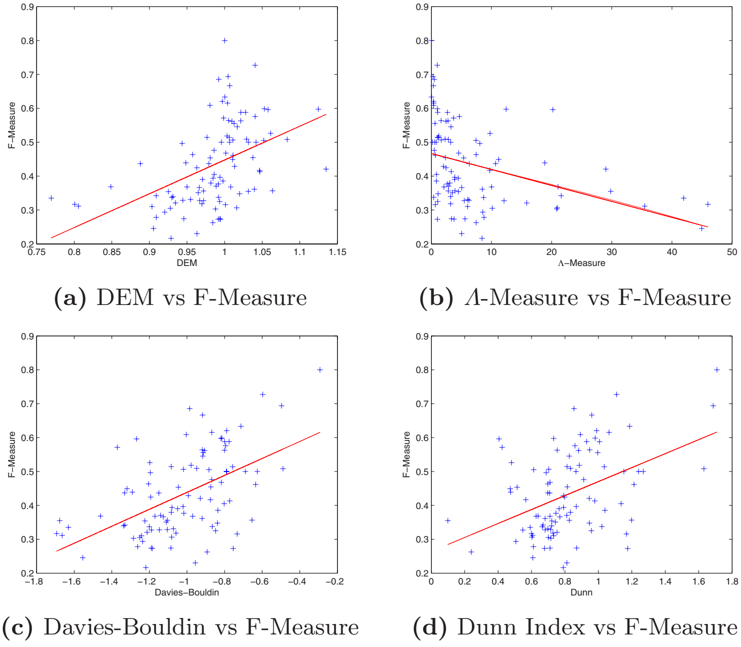
Figures 1, 2, 3 and 4 show the correlation results obtained for each corpus considered with the DEM, Λ-Measure, Davies-Bouldin and Dunn clustering validity measures. The x-axis corresponds to the different ICVM, whereas the y-axis corresponds to the F-Measure. In order to easily visualise the correlation between all the ICVM and F-Measure, we plotted the polynomial approximation of degree one. A desirable correlation would show points that start in the left corner



(a) DEM vs F-Measure **(b)** *Λ*-Measure vs F-Measure

(c) Davies-Bouldin vs F-Measure **(d)** Dunn Index vs F-Measure

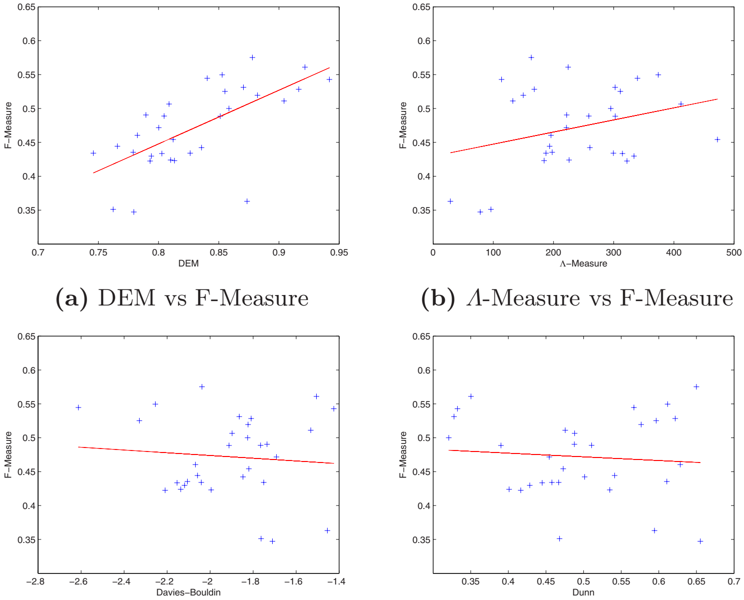
Fig. 1. Correlation of validity measures over the CICLing-2002 corpus



(a) DEM vs F-Measure **(b)** *Λ*-Measure vs F-Measure

(c) Davies-Bouldin vs F-Measure **(d)** Dunn Index vs F-Measure

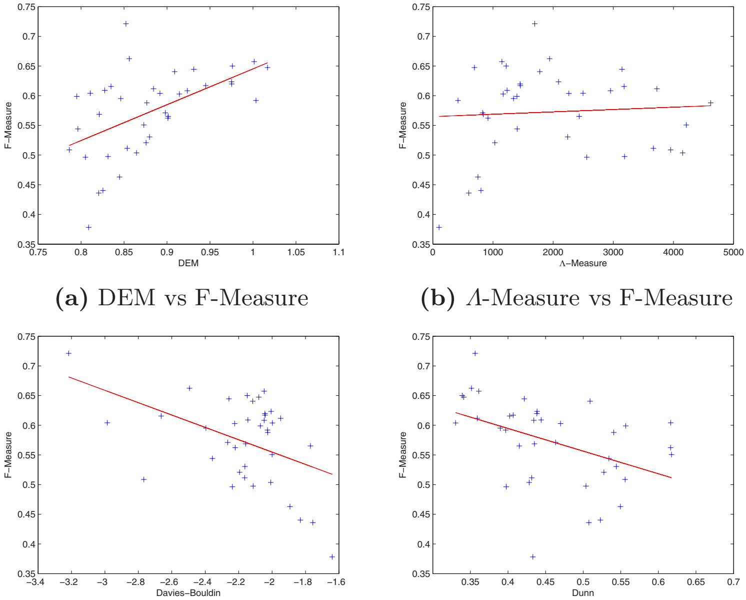
Fig. 2. Correlation of validity measures over the Semeval WSI corpus



(a) DEM vs F-Measure **(b)** *Λ*-Measure vs F-Measure

(c) Davies-Bouldin vs F-Measure **(d)** Dunn Index vs F-Measure

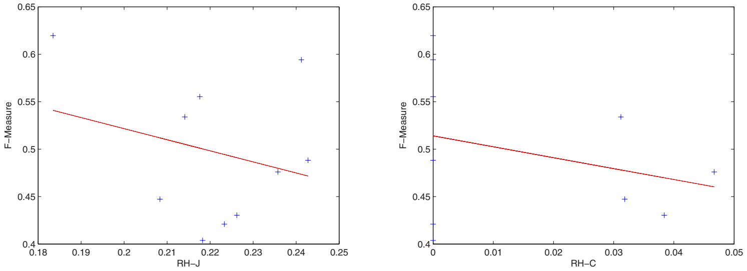
Fig. 3. Correlation of validity measures over the R8 test corpus



(a) DEM vs F-Measure **(b)** *Λ*-Measure vs F-Measure

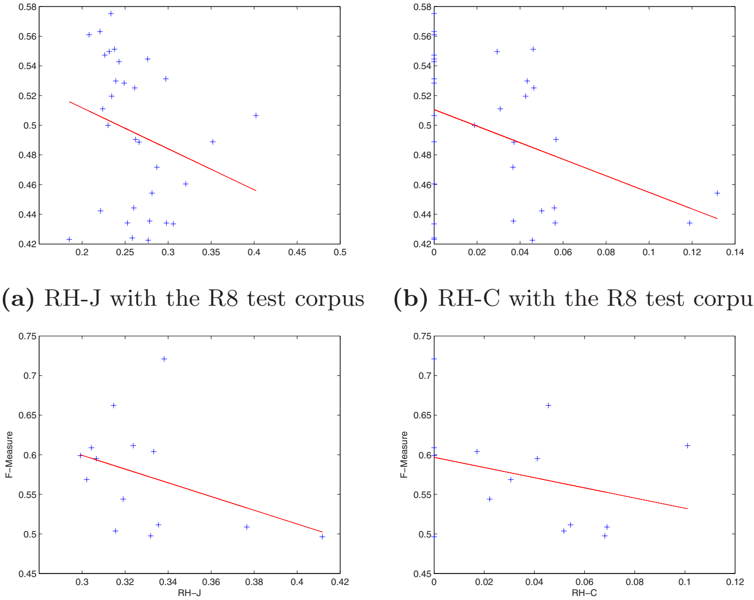
(c) Davies-Bouldin vs F-Measure **(d)** Dunn Index vs F-Measure

Fig. 4. Correlation of validity measures over the R8 train corpus



(a) RH-J **(b)** RH-C

Fig. 5. Evaluation of the CICLing-2002 corpus with the RH formulae based on the Jaccard coefficient and the cosine measure

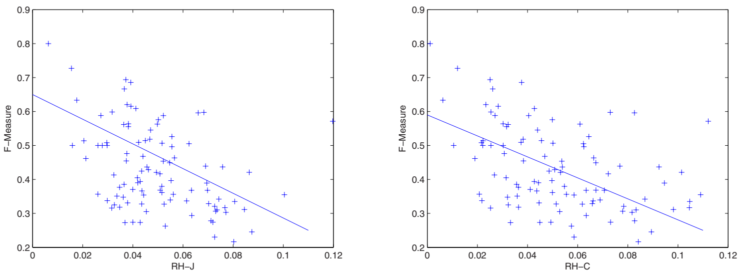


(a) RH-J with the R8 test corpus **(b)** RH-C with the R8 test corpus

(a) RH-J with the R8 train corpus **(b)** RH-C with the R8 train corpus

Fig. 6. Evaluation of the R8 test and train corpora with the RH formulae based on the Jaccard coefficient and the cosine measure

(low values of F-Measure) and grows monotonically (high values of F-Measure). In this sense, for a better readability we changed the sign of the Davies-Bouldin index, which is the only measure to be minimised and in this way the results are directly comparable. This modification was not done in Figures 5, 6 and 7, where we present the obtained results of the two introduced internal clustering



(a) RH-J

(b) RH-C

Fig. 7. Evaluation of the Semeval WSI collection with the RH formulae based on the Jaccard coefficient and the cosine measure

validity measures (RH-J and RH-C), since we wanted to emphasize the specific behaviour of these new measures.

We observed that DEM is the only measure analysed that keeps the expected direct correlation in all the corpora. This behaviour suggests a certain robustness of this measure. Specifically, when it is evaluated with the WSI Semeval corpora, it appears to have a lineal correlation with the F-Measure.

The Λ-Measure obtains an “acceptable” correlation with the *CICLing-2002* and R8 corpora. However it is remarkable that the correlation obtained with the WSI SemEval corpus is inverse. We conclude that this ICVM is not adequate in general for short texts. In order to be more precise with the degree of “acceptability”, as future work we aim to calculate some correlation index, such as the Spearman correlation coefficient which is a non-parametric (distribution-free) rank statistic which measures the strength of the associations between two variables [6]. One important finding is that if a clustering algorithm is designed in a way that attempts to optimise the Λ-Measure, then it will be negatively affected when using short-text corpora. The Davies-Bouldin index correlates very well with the F-Measure in the WSI SemEval collection, regular in the *CICLing-2002* corpus and quite bad in the R8 dataset. Finally, the Dunn measure behaves well with both, the *CICLing-2002* and WSI SemEval corpora, but it did not obtain a good correlation in the R8 dataset. We observed that the Davies-Bouldin and the Dunn indices have obtained similar results. With respect to the relative hardness, both, the one based on the Jaccard and the cosine similarity measures obtained good results in all the corpora.

From the corpora viewpoint, we may see that in the *CICLing-2002* corpus the ICVM measures obtain a good behaviour. In R8 all the results are consistent when evaluated in the test and train versions of this corpus; DEM, Λ-Measure and RH correlate very well with F-Measure, but Davies-Bouldin and Dunn obtain an inverse correlation. Future work should analyse the reason of these results. In WSI SemEval we obtained very good results for almost all ICVM (except Λ-Measure). The reader should pay attention that this collection consists of 100 corpora and, therefore, it makes sense to have obtained more stable results.

5 Conclusions and Future Work

The aim of this research work was to investigate whether the internal clustering validity measures may be used to improve the performance of clustering algorithms for short-text classification. In this paper we analysed different ICVMs with several short-text corpora.

Our findings indicate that the DEM and the RH measure are the ones that obtain the best results. However, it should be investigated whether the other ICVMs are related to specific kinds of corpora (for instance, narrow or wide domains). Thus, despite the corpora have very different characteristics it would be desirable to execute more experiments with other kind of domains, specifically to study, as we mentioned before, the narrow versus wide domain issue for short-text corpora. As further work we would also like to employ bio-inspired algorithms such as the DAntTree [4] to cluster short-text corpora. The main aim will be to investigate how to adapt the DAntTree algorithm to different internal clustering validity measures.

References

1. Agirre, E., Soroa, A.: Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In: Proc. of the SemEval Workshop, Prague, Czech Republic, The Association for Computational Linguistics, pp. 7–12 (2007)
2. Alexandrov, M., Gelbukh, A., Rosso, P.: An approach to clustering abstracts. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 8–13. Springer, Heidelberg (2005)
3. Bezdek, J.C., et al.: A geometric approach to cluster validity for normal mixtures. Soft Computing 1(4) (1997)
4. Ingaramo, D., Leguizamón, G., Errecalde, M.: Adaptive clustering with artificial ants. Journal of Computer Science & Technology 5(4), 264–271 (2005)
5. Karypis, G., Han, E.-H., Vipin, K.: Chameleon: Hierarchical clustering using dynamic modeling. Computer 32(8), 68–75 (1999)
6. Lehmann, E.L., D'Abrera, H.J.M.: Nonparametrics: Statistical Methods Based on Ranks. Prentice-Hall, Englewood Cliffs (1998)
7. Makagonov, P., Alexandrov, M., Gelbukh, A.: Clustering abstracts instead of full texts. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 129–135. Springer, Heidelberg (2004)
8. Montejo, A., Ureña, L.A.: Binary classifiers versus adaboost for labeling of digital documents. In: Procesamiento del Lenguaje Natural, pp. 319–326 (2006)
9. Pinto, D., Benedí, J.M., Rosso, P.: Clustering narrow-domain short texts by using the Kullback-Leibler distance. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 611–622. Springer, Heidelberg (2007)
10. Pinto, D., Rosso, P.: On the relative hardness of clustering corpora. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 155–161. Springer, Heidelberg (2007)
11. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering abstracts of scientific texts using the transition point technique. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 536–546. Springer, Heidelberg (2006)

12. Rose, T.G., Stevenson, M., Whitehead, M.: The Reuters Corpus volume 1 - from yesterday's news to tomorrow's language resources. In: Proc. of the 3rd International Conference on Language Resources and Evaluation - LREC 2002, pp. 827–832 (2002)
13. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
14. Shin, K., Han, S.Y.: Fast clustering algorithm for information organization. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 619–622. Springer, Heidelberg (2003)
15. Stein, B., Meyer, S., Wißbrock, F.: On cluster validity and the information need of users. In: Proceedings of the 3rd IASTED, pp. 216–221. ACTA Press (2003)