Proximity estimation and hardness of short-text corpora

Marcelo Luis Errecalde
and Diego Ingaramo
Development and Research Lab. in Computational Intelligence
Universidad Nacional de San Luis, Argentina
Email: {merreca,daingara}@unsl.edu.ar

Paolo Rosso
Natural Language Engineering Lab.
Department of Information Systems and Computation
Universidad Politécnica de Valencia, Spain
Email: prosso@dsic.upv.es

***Abstract*—In this work, we investigate the relative hardness of short-text corpora in clustering problems and how this hardness relates to traditional similarity measures. Our approach basically attempts to establish a connection between the hardness of a corpus and the precision level exhibited by similarity measures, according to the results obtained with different cluster validity measures on the "ideal" clustering of each corpus. Moreover, we also propose a new validity measure, named *contiguity error* that allowed us to observe this connection in a consistent way in all the collections considered.**

I. Introduction

Studies on clustering problems often assume that an accurate proximity measure between objects to be clustered is available and they only pay attention to the principles behind the functioning of clustering algorithms. This assumption is usually valid in some simple domains and, in particular, in "geometric" domains where objects correspond to data points in the Euclidean space and the proximity measure is based on the Euclidean distance. However, when clustering tasks involve documents, different aspects can negatively affect the proximity estimation between documents.

A very popular document representation scheme used in these cases is the Vector Space Model (VSM) [1] where documents are represented as vectors of terms and the underlying metaphor consists in using spatial proximity for semantic proximity. The usual approach consists in estimating the similarity of two documents using the cosine of the angle between the corresponding high dimension vectors indexed by the terms in the corpus. This is an effective and conceptually simple approach, but it is nevertheless an oversimplification. A strong assumption in these cases is that the terms are independent, i.e., the dimensions of the input space are orthogonal. However, in text applications, the input space is usually non-orthogonal due to problems of synonymy and polysemy. Therefore, the similarity measured by cosine or inner product based on Euclideans distance cannot exactly describe the relationship between documents [2].

The similarity measurement is an important issue since it affects different aspects related to the clustering process:

- *Cluster validation*: most of internal indexes used in cluster validity are based on similarity (or dissimilarity) measures. See section II-A1 for a more detailed discussion.
- *Clustering as an optimization problem*: several approaches visualize clustering as a problem where a given arbitrary objective function must be optimized. In these cases, any unsupervised measure of cluster validity can be used as an objective function and these functions are usually based on a similarity measure [3].
- *Clustering algorithms' robustness*: a reasonable estimation of how accurate a similarity measure is can be useful for determining those cases where a poor clustering result can be ascribed to a deficient similarity estimation and those cases where a bad performance is caused by limitations of the clustering algorithms.

This work has been partially supported by the MCyT TIN2006-15265-C06-04 project, the ANPCyT and the Universidad Nacional de San Luis.

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 [4].

In order to obtain a better understanding of the complexity in clustering short-text corpora, a deeper analysis of the hardness of this kind of corpora is required. Specifically, we are interested in answering the following questions:

- it is usually assumed that short text corpora are harder to deal with than traditional corpora, but *how*?
- *how accurate* traditional similarity measures in these cases are?
- to what extent are both issues related?

To answer these questions we do not use any clustering method at all. We propose an approach instead where these aspects are inferred from a meticulous analysis based on different validity cluster measures when evaluated on the "ideal clustering" of each corpus. We consider two different very short-text corpora which differ in the overlapping degree of their vocabularies. Results are also compared with a corpus which contains longer documents on well differentiated topics. In a nutshell, we want consider situations where these measures adequately express the conceptual proximity of documents and other cases where the similarity measures exhibit different levels of noise (error).

The remainder of the paper is organized as follows. Section II presents our view about the hardness problem and its relation with similarity measures. In Section III some general features of the corpora used in the experiments are presented. Section IV analyzes the relative hardness of these corpora considering the criteria stated in section II. Finally, some general conclusions are drawn and possible future work is discussed.

II. Relative Hardness of short-text corpora

Relative hardness is the expression introduced in [5] to refer to the relative difficulty of different Reuters-21578 subsets for supervised categorization tasks. In this case, a reliable estimation of the relative difficulty of these subsets is established from the results obtained in a variety of experimental text categorization contexts.

In [4] on other hand, the hardness concept is considered in a more general context and the authors propose a formula for quantify this hardness based on the vocabulary overlapping among the categories of a corpus. They experiment with all the possible subcorpus of

different standard collections using the MajorClust algorithm [6] for clustering each resulting subcorpus. The results show an interesting correlation level between the values obtained with their formula and the popular F-measure.

In the present work we will take a different perspective and we will consider the relative hardness of a corpus respect to the difficulty level that it presents for establishing an *accurate similarity measure* between its documents. Our approach will be tested with three corpora which are assumed to have different difficulty level: we are interested in detecting the relative hardness of short-text corpora and narrow domain short-text corpora respect to other more standard corpora. This approach implies that a criterion for estimating how accurate a similarity measure is, must be defined. This criterion is introduced in the following subsection.

A. Evaluating similarity measures

Of central importance in attempting to identify clusters of documents is knowledge of how "close" documents are to each other, or how far apart they are. This quantitative measure of closeness is commonly referred in this context as *dissimilarity*, *distance* or *similarity*, with a general term being *proximity* [7]. Two documents are 'close' when their dissimilarity or distance is small or their similarity large. From now on we will assume that a similarity measure sim is available. The attempt is to quantify the relatedness degree of two documents according to a given criterion. More formally, let $\mathcal{D}$ be an arbitrary collection of documents, a similarity measure $sim : \mathcal{D} \times \mathcal{D} \rightarrow R$ is a mapping such that $\forall d_1, d_2 \in \mathcal{D}$, $sim(d_1, d_2)$ quantifies how similar the documents d_1 and d_2 are.

Criteria for determining if two documents will be considered similar can significatively vary depending on the particular clustering problem at hand. Therefore, the relevant information required to compute an adequate similarity measure can also significatively vary in each case. For example, in traditional document clustering problems we need to express the similarity between documents with respect to topics. Therefore, the vocabulary of the documents will have a great influence in the similarity measurement although some words that do not contain topical information (e.g. stop-words) are usually filtered. On the other hand, when documents have to be clustered by author, stylographic features are usually more relevant and, in contrast to the previous case, to use frequencies of stop-words can be informative.

Despite the relevance of similarity measures for clustering tasks, to establish how accurate a similarity measure is for a particular problem is not a simple task. In this case, information about the correct similarity measurement between any pair of documents in the collection should be available. However, in clustering documents problems, people are not usually able to provide this kind of information. An expert in a particular domain will have no problem for deciding whether two documents belong to the same category or not. These capabilities have been useful, for example, for helping to (semi)-supervised clustering algorithms to achieve better results [8]. Likewise, in some cases, this expert can also be able to determine if a document d_k is more similar to a document d_l than to a third document d_m. However, this *qualitative* information about similarity between documents is not very useful for determining how precise a *quantitative* similarity measure is.

As can be noted, the task of accurately estimating how correct (or incorrect) a similarity measure is, can constitute a problem as complex as the clustering problem itself. However, there are situations where we can affirm (at least with a considerable confidence level) whether a particular similarity measure is working or not. When a gold standard is available, it is possible to do an analysis of this kind. An alternative consists of using the same internal validity measures used to evaluate the results of clustering and to apply them directly to the "correct" grouping defined by the human expert. This idea is reasonable if we think that these measures are mainly based on the similarity measure. If these measures are not able to detect any interesting structural property when applied to the "ideal clustering", this fact can be considered enough evidence that the similarity measure is not a good indicator of the semantic proximity between documents.

An obvious question arises with this approach: which validity measure should be used? There is not a unique answer to this question since it depends on which are the properties we expect the clustering satisfies and, therefore, what result we will consider a correct clustering. In other words, when an internal validity measure is selected for evaluating a clustering, we already have in mind some type of correct clustering . For example, some internal validity measures only attempt to determine to what extent the clusters are "well separated". In other cases, attention is paid to express the cohesion degree between documents belonging to the same groups. However, traditional measures usually combine both aspects and give a high score to clustering where documents in the same group exhibit high cohesion and elements from different groups are well separated. Other measures are *density-based* and they rely on the idea that clusters are regions of high density separated by regions of low density. Another valid criterion to be expressed in a validity measure is *contiguity*, which considers if each point is closer to at least one point in its cluster than to any point in another cluster. For a simple and comprehensive description of different types of clusters and measures for evaluating them see [9].

We address this problem avoiding establish a commitment with a particular validity measure and considering a representative group of measures instead. In Section IV some situations will be considered where a set of popular and representative validity measures will be applied to the correct clustering and where we can conclude (with a considerable level of confidence) that the similarity measure is adequately working. Other situations will also be considered, where a poor performance of the majority of validity measures is significative evidence that some level of noise (error) is probably present in the similarity measure computation. Next, some considerations on different validity measures used in this work are presented.

1) Validity Measures: Cluster validity is a measure of goodness for results obtained by clustering algorithms. There exist two types of cluster validity measures, namely, *external* and *internal*. The difference relies, respectively, on the use or not of a pre-specified structure of the data which is usually imposed by an expert.

External validity measures include the well-known F-Measure and the Entropy. Examples of traditional internal validity measures are the *Dunn Index Family*, the *Davies-Bouldin Index* and the *Silhouette Coefficient* but other more recent proposal like the Λ-*Measure* and the *Density Expected Measure* $\overline{\rho}$ are described in [10]. Space constraints make impossible an adequate description of these measures but more comprehensive explanations can be obtained in [10], [9]. An exception is the *contiguity error* measure described below that we introduce in this work for detecting possible contiguity errors of the similarity measure.

The Contiguity Error: we define this measure in order to have another perspective about the noise (error) exhibited by the similarity measure. Basically, the *Contiguity Error* (CE) is an external validity measure because it requires to know a reference categorization for determining which cluster is assigned by the expert to each

document. Intuitively CE counts how many contiguity errors the similarity measure produces respect to the clustering specified by the expert. In other words, CE computes the number of documents which have as nearer neighbour (according to the similarity measure sim) a document belonging to a different cluster (according to the expert's categorization). Formally, let D denote the set of documents under consideration, and let $\mathcal{C}^* = \{C_1^*, \ldots, C_l^*\}$ be the reference categorization. Let $cla^* : D \rightarrow \mathcal{C}^*$ be a mapping such that if $cla(d_i) = C_j^*$, C_j^* is the class assigned to document d_i according to the expert's classification $\mathcal{C}^*$. If $sim : \mathcal{D} \times \mathcal{D} \rightarrow R$ is a similarity measure between documents, the *Contiguity Error* CE of sim respect to $\mathcal{C}^*$, is defined as

$$CE(sim) = \sum_{d \in D} nco(d)$$

where

$$nco(d) = \begin{cases} 0 & \texttt{if } ms(d) = d_k \wedge cla(d) = cla(d_k), \\ 1 & \texttt{other cases.} \end{cases} \quad (1)$$

and $ms : \mathcal{D} \rightarrow \mathcal{D}$ gives the *most similar* neighbour to a document, i.e., if $ms(d) = d_i \Rightarrow \forall d_k \in \mathcal{D}, sim(d, d_i) \geq sim(d, d_k)$.

When working with collections with different number of documents it can be more informative computing the contiguity error *percentage* produced by sim ($CEP(sim)$) respect to the total number of documents:

$$CEP(sim) = \frac{CE(sim)}{|\mathcal{D}|}$$

III. Data Sets

The complexity of clustering problems with short-text corpora demands a meticulous analysis of the features of each collection used in the experiments. For this reason, we will focus on specific characteristics of the collections such as document lengths and its closeness respect to the topics considered in these documents. We attempt with this decision to avoid introducing other factors that can make incomparable the results.

With the exception of CICling-2002 collection which has already been used in previous works [11], [12], [13], the remaining two corpora were artificially generated with the goal of obtaining corpora with different levels of complexity respect to the length of documents and vocabulary overlapping. Our intention was that in each corpora the similarity measure has different levels of complexity for detecting the conceptual proximity between documents. However, other features such as the number of groups and number of documents per group were maintained the same for all collections in order to obtain comparable results.

It could be argued that our preliminary analysis is limited to small size collections. However, we believe that short-text clustering in general and clustering of narrow domain abstracts in particular, demand a detailed understanding of each collection that would be difficult to achieve with large size standard corpora.

In the following subsections, a general description of two collections used in this work is presented. These collections are introduced in increasing order of complexity. We begin with the Micro4News corpus, a collection of medium-length documents about well differentiated topics (low complexity). Then, the EasyAbstracts corpus with short-length documents (scientific abstracts) and well differentiated topics is presented (medium complexity corpus). We created these two new collections with similar general characteristics (number of groups and number of documents per group). [1] The CICling-2002 corpus with relatively high complexity was also used in our work. This collection is considered to be harder to cluster than the previous corpora since its documents are narrow domain abstracts (see [13] for a more detailed description of the corpus).

A. The Micro4News Corpus

This first collection was constructed with medium-length documents that correspond to four very different topics. Consequently, in this case it is supposed that the similarity measure will not have any problem in determining if two documents are semantically related. Its documents are significatively larger than CICling-2002 and talk about well differentiated topics. We select documents belonging to four very different groups of the popular 20Newsgroups corpus [15]: 1) *sci.med*, 2) *soc.religion.christian*, 3) *rec.autos* and 4) *comp.os.ms-windows.misc*. For each topic, the largest documents were selected. Thus, we ensure that the average length of its documents were seven times (or more) the length of abstracts of the remaining two corpora.

B. The EasyAbstracts Corpus

This collection can be considered harder than the previous one because its documents are scientific abstracts (same characteristic as CICling-2002) and in consequence are short documents. It differs from CICling-2002 respect to the overlapping degree of the documents' vocabulary. EasyAbstracts documents also refer to a shared thematic (*intelligent systems*) but its groups are not so closely related as the CICling-2002 groups are. EasyAbstracts was constructed with abstracts publicly available on Internet that correspond to articles of four international journals in the following fields: 1) *Machine Learning*, 2) *Heuristics in Optimization*, 3) *Automated reasoning* and 4) *Autonomous intelligent agents*. It is possible to select abstracts for these disciplines in a way that two abstracts of two different categories are not related at all. However, some degree of complexity can be introduced if abstracts of articles related to two or more EasyAbstracts's categories are used. [2] We included in the EasyAbstract corpus a few documents with these last features in order to increase the complexity respect to the Micro4News corpus. Nevertheless, the majority of documents in this collection clearly belong to a single group. This last fact allows us to assume that a similarity measure should not have any problem in representing the proximity among documents compared with the complexity of CICling2002 corpus.

IV. Similarity estimation and Corpora Analysis

There are two main factors that usually impact on a similarity measure between documents: the document representation and the procedure used for computing the similarity between documents with this representation. One of the most widely used model for document representation is VSM which has associated a family of weighting schemes that we will refer as the "SMART codifications". Here, vector (document) similarity is usually measured by the *cosine* measure but other similarity measures derived from the Euclidean distance can also be used with this representation. Another popular document representation approach is the *set model* which considers a document as a set whose elements are the document's terms. In this case, proximity between documents is often quantified by set

[1] A detailed description of the distribution and features of this two corpora is available in [14] where you can also access the corpora for research purposes.

[2] For instance, abstracts which refer to *learning intelligent agents* or agents with high level reasoning capabilities.

intersection ratios being the *Jaccard coefficient* one of the most popular scheme for measuring set similarity.

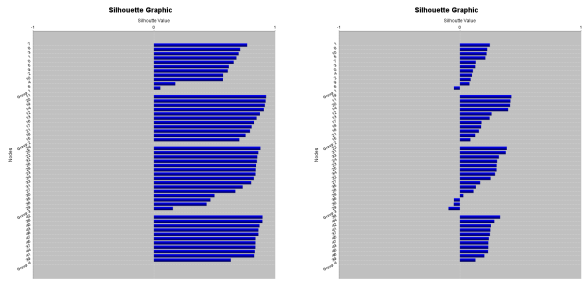
In our work, we used the Jaccard coefficient and the SMART [1] system conventional code scheme with the cosine similarity measure. In the SMART system, each codification is composed by three letters: the first two letters refer, respectively, to the TF (Term Frequency) and IDF (Inverse Document Frequency) components, whereas the third one ($NORM$) indicates whether normalization is employed or not. Taking into account standard SMART nomenclature, we will consider five different alternatives for the TF component: n (natural), b (binary), l (logarithm), m (max-norm) and a (aug-norm); two alternatives for the IDF component (n and t) and two alternatives for normalization: n (no normalization) and c (cosine). In this way, a codification ntc will refer to the popular scheme where the weight for the i-th component of the vector for the document d is computed as $tf_{d,i} \times \log(\frac{N}{df_i})$ and then cosine normalization is applied. Here, N denotes the number of documents in the collection, $tf_{d,i}$ is the term frequency of the i-th term in the document d and df_i refers to the document frequency of i-th term over the collection (see [16] for a more detailed explanation). With this representation scheme we can generate 20 different codifications but we will only report results with the 10 normalized codifications ("**c" codifications) because codifications without normalization give equivalent results when cosine similarity is used as proximity measure.

Micro4News Corpus: As it was explained in section II-A, our performance analysis of the similarity measure (and, therefore, the relative hardness) will be based on the results delivered by a representative set of validity measures on the "correct" clustering. In particular we will focus on values obtained with the following validity indexes: Contiguity Error (CE), Density Expected Measure (DEM), Davies- Bouldin Index (DB), Dunn Index (Dunn) and Global Silhouette (GS). It is important to observe that large values of these measures correspond to a good cluster with the exception of DB Index where small values are considered goods. Table I presents results of the validity measures obtained with: a) SMART codifications and cosine similarity and, b) Jaccard Coefficient (denoted Jac).

Codif.	CE	DEM	DB	Dunn	GS
atc	0	0.9	1.64	0.76	0.46
btc	0	0.9	1.64	0.76	0.46
mtc	0	1.07	1.33	0.76	0.73
ntc	0	1.07	1.34	0.74	0.73
Jac	0	0.78	2.10	0.50	0.2
anc	1	0.77	2.48	0.85	0.16
ltc	1	0.92	1.59	0.77	0.50
bnc	1	0.77	2.45	0.85	0.17
lnc	1	0.78	2.52	0.87	0.14
mnc	10	0.82	2.89	0.75	0.02
nnc	10	0.82	3.38	0.74	0.02

TABLE I
MICRO4NEWS: VALUES OF VALIDITY MEASURES

Here, it can be observed that traditional ntc codification with cosine similarity gives very good values on all validity measures except the Dunn Index. As an example, if we consider that $CE = 0$, that means that the similarity measure does not commit any contiguity error, i.e., for each document, its most similar document belongs to the same group. There are other four SMART codifications with CE values equal to 0, but ntc shows the better values for DEM (1.07), GS (0.73) and the second best value of DB (1.34). In that sense, should be noted that the mtc codification is another valid candidate to be selected as the "best" codification.



(Cosine) **(Jaccard)**

Fig. 1. Micro4News: Silhouette graphics.

An interesting result arise when *Jaccard* is used as similarity measure. In this case, we obtain the worst results for the DEM and $Dunn$ metrics. The Global Silhouette value is also very low (0.2), and this fact would indicate that a cluster structure is absent considering this measure. However, Jaccard gets a good score respect to the CE criterion.

In Figure 1 silhouette graphics are shown for the best SMART codification with similarity cosine (ntc) and for Jaccard similarity. In the first case, each document shows an evident membership degree to its group but results with Jaccard are not so good.

EasyAbstracts Corpus: in this collection, the values obtained with the validity measures are not so good as in the previous collection as can be observed in Table II. If we consider the CE measure, we can see that the best results (ntc and mtc codifications) give at least four contiguity errors and for the worst cases (mnc and nnc codifications) the contiguity errors are five times the values of the best results, i.e., we have 20 documents with contiguity errors.

Codif.	CE	DEM	DB	Dunn	GS
mtc	4	0.93	1.57	0.71	0.47
ntc	4	0.93	1.57	0.71	0.47
ltc	5	0.89	1.7	0.71	0.33
atc	5	0.88	1.72	0.71	0.31
btc	6	0.88	1.74	0.71	0.28
lnc	11	0.73	3.57	0.86	0.07
anc	11	0.72	3.49	0.85	0.07
Jac	13	0.74	2.15	0.5	0.08
bnc	15	0.72	3.28	0.82	0.07
mnc	20	0.75	4.91	0.87	0.02
nnc	20	0.75	4.91	0.87	0.02

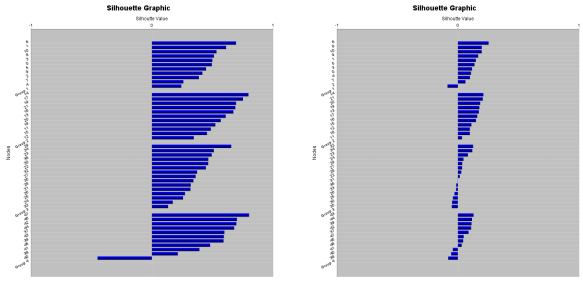
TABLE II
EASYABSTRACT: VALUES OF VALIDITY MEASURES

In this collection, the ntc and mtc codifications with cosine similarity again yield the best results for the majority of the validity measures. Thus, we can see that the best values for DEM (0.93), DB (1.57) and GS (0.47) are obtained with these representations.

With *Jaccard* similarity, low values are obtained for DEM (0.74), Dunn (0.5) and GS (0.08), and consequently these values could be interpreted as a lack of cluster structure in the groups when this measure is used. Moreover, a CE value of 13 obtained with this collection indicates a high number of contiguity errors.

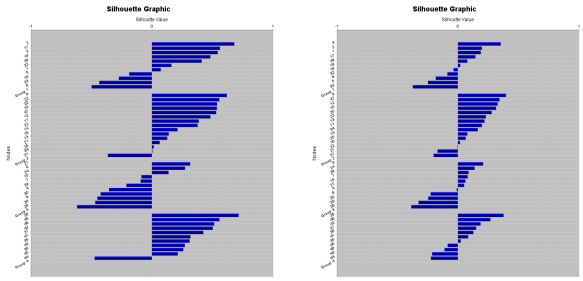
Figure 2 shows that according to the silhouette index, in this collection the membership degree of the documents to their respective clusters is not so high as the achieved with the Micro4News collection. The results also show that cosine similarity with ntc codification clearly outperforms Jaccard coefficient.

CICLing-2002 Corpus: without any doubt is in this collection



(Cosine) **(Jaccard)**

Fig. 2. EasyAbstracts: Silhouette graphics.



(Cosine) **(Jaccard)**

Fig. 3. CICLing2002: Silhouette graphics.

where we can expect to observe the most unstable results in the validity measures. This can be appreciated in the Table III.

Codif.	CE	DEM	DB	Dunn	GS
mnc	16	0.8	2.21	0.79	0.15
nnc	16	0.8	2.21	0.79	0.15
btc	18	0.84	1.82	0.74	0.07
anc	21	0.76	2.45	0.8	0.07
Jac	22	0.79	2.28	0.53	0.05
atc	22	0.85	1.8	0.74	0.1
bnc	22	0.75	2.51	0.8	0.04
ltc	23	0.85	1.8	0.74	0.1
lnc	23	0.76	2.45	0.8	0.08
mtc	23	0.87	1.76	0.74	0.15
ntc	23	0.87	1.76	0.74	0.15

TABLE III
CICLING2002: VALUES OF VALIDITY MEASURES

If we consider the ntc codification with cosine similarity, we can see that this scheme obtains the best values in the following indexes: DEM, DB and GS confirming the tendency observed in the previous collections. However, its CE value (23) is one of the worst values for this collection. We can also appreciate that uniform and relatively bad values of DEM, CE and GS are obtained for all codifications considered. This fact is indicative of the difficulty that validity measures have for capturing the structural properties of the clustering using these similarity measures. This lack of cluster structure is clearly appreciated in the silhouette graphics of Figure 3.

V. CONCLUSIONS

Results obtained in the previous section are indicative that our approach can be useful for determining the hardness of corpora used as testbed in clustering of short-text corpora. We can also conclude that traditional schemes for computing similarity measures can be used with short-text corpora with well differentiated topics but its performance can be seriously affected in narrow domain short-text corpora. This fact suggests that in this kind of domains significative work is required for obtaining more adequate similarity measures. A good starting point would be to test with our approach some recent and more elaborated schemes [17], [18] .

In collections such as CiCling-2002, all validity measures used in this work had serious problems for expressing structural properties of the clustering. However, Silhouette Global, Density Expected Measure and Contiguity Error exhibit an interesting consistency level between all the collections considered and seem to be the most informative for detecting which representation and similarity scheme is the most adequate for each corpus.

Our study also aims to identify those cases where a poor clustering result can be ascribed to a deficient similarity estimation and those cases where a bad performance is caused by limitations of the clustering algorithms. With respect to this point, we are currently testing six different representative clustering algorithms on different short-text corpora for observing their robustness to the different error (noise) level exhibited for the similarity measures.

REFERENCES

[1] G. Salton, *The Smart Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall, 1971.
[2] N. Liu, B. Zhang, J. Yan, Q. Yang, S. Yan, Z. Chen, F. Bai, and W.-Y. Ma, "Learning similarity measures in non-orthogonal space," in *CIKM '04*. ACM, 2004, pp. 334–341.
[3] Y. Zhao and G. Karypis, "Empirical and theoretical comparisons of selected criterion functions for document clustering," *Machine Learning*, vol. 55, no. 3, pp. 311–331, 2004.
[4] D. Pinto and P. Rosso, "On the relative hardness of clustering corpora," in *Proc. of the Text, Speech and Dialogue 2007 Conference - TSD07*, ser. LNAI, vol. 4629. Springer-Verlag, 2007, pp. 155–161.
[5] F. Debole and F. Sebastiani, "An analysis of the relative hardness of Reuters-21578 subsets," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 6, pp. 584–596, 2004.
[6] B. Stein and O. Niggemann, "On the Nature of Structure and its Identification," in *Proc. of the 25th International Workshop on Graph Theoretic Concepts in Computer Science - WG99*, ser. Lecture Notes in Computer Science, vol. 1665. Springer-Verlag, 1999, pp. 122–134.
[7] B. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, 2001.
[8] E. Xing, A. Y. Ng, M. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side information," in *NIPS2002*, 2002, pp. 521–528.
[9] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 2006.
[10] B. Stein, S. Meyer zu Eissen, and F. Wißbrock, "On cluster validity and the information need of users," in *Proceedings of the 3rd IASTED*. ACTA Press, 2003, pp. 216–221.
[11] P. Makagonov, M. Alexandrov, and A. Gelbukh, "Clustering abstracts instead of full texts," in *Proc. of the TSD-2004 Conference*, ser. LNAI, vol. 3206. Springer-Verlag, 2004, pp. 129–135.
[12] M. Alexandrov, A. Gelbukh, and P. Rosso, "An approach to clustering abstracts," in *Proc. of the 10th Int. NLDB-05 Conference*, ser. Lecture Notes in Computer Science, vol. 3513. Springer-Verlag, 2005, pp. 8–13.
[13] D. Pinto, J. M. Benedí, and P. Rosso, "Clustering narrow-domain short texts by using the Kullback-Leibler distance," in *Proc. of the CICLing 2007 Conference*, ser. Lecture Notes in Computer Science, vol. 4394. Springer-Verlag, 2007, pp. 611–622.
[14] M. Errecalde and D. Ingaramo, "Short-text corpora for clustering evaluation," LIDIC, Tech. Rep., 2008. [Online]. Available: http://www.dirinfo.unsl.edu.ar/∼ia/resources/shortexts.pdf
[15] K. Lang, "20 newsgroups, the original data set," 1993, http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html.
[16] C. D. Manning and H. Schtze, *1999*, F. of Statistical Natural Language Processing, Ed.
[17] B. Stein, S. Meyer zu Eissen, and M. Potthast, "Syntax versus semantics: Analysis of enriched vector space models," in *Third International Workshop on Text-Based Information Retrieval (TIR 06)*, B. Stein and O. Kao, Eds. University of Trento, Italy, August 2006, pp. 47–52.
[18] A. Hotho, S. Staab, and G. Stumme, "Wordnet improves text document clustering," in *Proc. of the Semantic Web Workshop at SIGIR-2003*, 2003.