# On the Relative Hardness of Clustering Corpora[*]

David Pinto[1,2] and Paolo Rosso[1]

[1] Department of Information Systems and Computation,
Polytechnic University of Valencia, Spain
Faculty of Computer Science
[2] B. Autonomous University of Puebla, Mexico
{dpinto, prosso}@dsic.upv.es

**Abstract.** Clustering is often considered the most important unsupervised learning problem and several clustering algorithms have been proposed over the years. Many of these algorithms have been tested on classical clustering corpora such as Reuters and 20 Newsgroups in order to determine their quality. However, up to now the relative hardness of those corpora has not been determined. The relative clustering hardness of a given corpus may be of high interest, since it would help to determine whether the usual corpora used to benchmark the clustering algorithms are hard enough. Moreover, if it is possible to find a set of features involved in the hardness of the clustering task itself, specific clustering techniques may be used instead of general ones in order to improve the quality of the obtained clusters. In this paper, we are presenting a study of the specific feature of the vocabulary overlapping among documents of a given corpus. Our preliminary experiments were carried out on three different corpora: the train and test version of the R8 subset of the Reuters collection and a reduced version of the 20 Newsgroups (Mini20Newsgroups). We figured out that a possible relation between the vocabulary overlapping and the F-Measure may be introduced.

## 1 Introduction

Clustering deals with finding a structure in a collection of unlabeled data [2]. When dealing with raw text corpora, the discovering of the most appropiate features can help on the selection of methods and techniques for determining the possible intrinsic grouping in those sets of unlabeled data. Therefore, this study would be of high benefit. As far as we know, research works in this field nearly have not been carried out in literature. We found just one attempt for determining the relative hardness of the Reuters-21578[1] clustering collection [1], but this research work neither derived formulae for determining the hardness of these corpora nor the possible set of features that are involved in the clustering hardness. A related work which could be considered in order to observe the hardness of a given corpus (with respect to a specific clustering algorithm) is partially

[*] The term 'hardness' is employed like in [1] where this term was introduced to analyse the relative hardness of the Reuters corpora.

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

V. Matousek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 155–161, 2007.

presented in [3] and [4]. In these research works, the author discusses internal clustering quality measures, such as the one from the Dunn Index family, which showed to perform well in the experiments presented by Bezdek et al. in [5,6], among others.

Reuters-21578 (now Reuters RCV1 and RCV2) and 20 Newsgroups[2] are well-known collections which have been used for benchmarking clustering algorithms. However, the fact that several clustering methods may obtain bad results over those corpora does not necessarily imply that they are difficult to be clustered. Further investigation needs to be done in order to determine whether the current clustering corpora are easy clustering instances or not.

We are interested in investigating two aspects: a set of possible features hypothetically related with the hardness of the clustering task, as well as the definition of a formula for the easy evaluation of the relative hardness of a given clustering corpus. We empirically know that at least three components are involved: (i) the size of the clustering texts, (ii) the broadness of the corpora domain and, (iii) whether the documents are single or multi categorized. In the our preliminary experiments, we have investigated the possible relationship between the vocabulary overlapping of a given text corpus with its F-Measure obtained with the MajorClust clustering algorithm [7].

The rest of this paper is structured as follows. In Section 2 we briefly describe the main characteristics of the corpora used in our preliminary experiments. In Section 3 we introduce the used formula and the employed approach to split the corpus in order to calculate the relative hardness for all the possible combinations of two or more categories. Section 5 shows the experimental results we obtained. Finally, conclusions are drawn and the necessary further work to be done is discussed.

## 2 Datasets

The preliminary experiments were carried out by using three different corpora: the R8 version of the Reuters collection (train and test) and, partially, a reduced version of the 20 Newsgroups named "Mini20Newsgroups". We have pre-processed each corpus eliminating punctuation symbols, stopwords and, thereafter, applying the Porter stemmer. The characteristics of each corpus after the pre-processing are given in Table 1.

**Table 1.** Characteristics of Reuters-R8 and Mini20Newsgroups

| | **R8-Train** | **R8-Test** | **Mini20Newsgroups** |
|---|---|---|---|
| **Size** | ≈2,500 KBytes | ≈900 KBytes | ≈1,900 KBytes |
| **Documents** | 5,839 | 2,319 | 2,000 |
| **Categories** | 8 | 8 | 20 |

## 3 Calculating the Relative Hardness of a Corpus

In order to determine the Relative Hardness (RH) of a given corpus, we have considered the vocabulary overlapping among the texts of the corpus. In our experiments, we

[2] `http://people.csail.mit.edu/jrennie/20Newsgroups/`

have used the well-known Jaccard coefficient for calculating the overlapping. We considered all the possible combinations of more than two categories from the corpus and for each of them we calculated its RH. For instance, for a given corpus of $n$ categories, $2^n - (n+1)$ possible subcorpora will be obtained: e.g. for the R8 (eight categories) we obtained 247 subsets.

Thereafter, we calculated their RHs as follows: given a corpus $C_i$ made up of $n$ categories (CAT), the RH of $C_i = \{CAT_1, CAT_2, ..., CAT_n\}$ is:

$$RH(C_i) = \frac{1}{n(n-1)/2} \times \sum_{j,k=1;j<k}^{n} Similarity(CAT_j, CAT_k), \tag{1}$$

where the similarity among categories is obtained by using the Jaccard coefficient in order to determine their overlapping (see Eq. (2)). However, more sophisticated measures also could be used, such as the one presented in [8] in the plagiarism degree calculation framework.

$$Similarity(CAT_j, CAT_k) = \frac{|CAT_j \bigcap CAT_k|}{|CAT_j \bigcup CAT_k|} \tag{2}$$

In the above formula we have considered each category $j$ as the “document” obtained by concatenating all the documents belonging to the category $j$.

## 4 Clustering the Datasets

In order to evaluate the relative hardness formula used in the experiments, we have carried out an unsupervised clustering of all the documents of each subcorpus obtained for each dataset. We have chosen the MajorClust clustering algorithm [7] due to its peculiarity of taking into account both, the inside and outside similarities among the clusters obtained during its execution. In order to keep independent the validation with respect to RH, we have used the tf-idf formula for calculating the input similarity matrix for MajorClust. Each evaluation was performed with the F-Measure formula which is calculated as follows: given a set of clusters $\{G_1, \ldots, G_m\}$ and a set of classes $\{C_1, \ldots, C_n\}$, the $F$-measure between a cluster $i$ and a class $j$ is given by the following formula.

$$F_{ij} = \frac{2 \cdot P_{ij} \cdot R_{ij}}{P_{ij} + R_{ij}}, \tag{3}$$

where $1 \leq i \leq m$, $1 \leq j \leq n$. $P_{ij}$ and $R_{ij}$ are defined as follows:

$$P_{ij} = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts from cluster } i}, \tag{4}$$

and

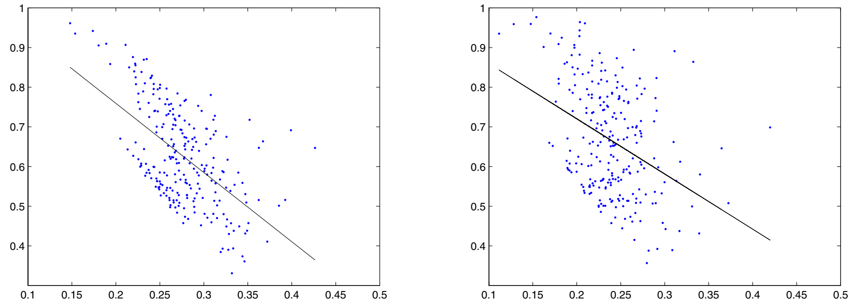
$$R_{ij} = \frac{\text{Number of texts from cluster } i \text{ in class } j}{\text{Number of texts in class } j}. \tag{5}$$

The global performance of a clustering method is calculated by using the values of $F_{ij}$, the cardinality of the set of clusters obtained, and normalising by the total number of documents in the collection ($|D|$). The obtained measure is named $F$-measure and it is shown in Equation (6).

$$F = \sum_{1 \leq i \leq m} \frac{|G_i|}{|D|} \max_{1 \leq j \leq n} F_{ij}. \quad (6)$$

## 5 Correlation Between Relative Hardness and F-Measure

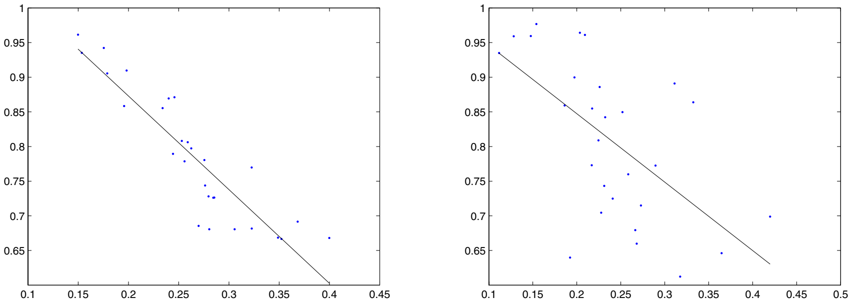
Our preliminary experiments were carried out on the train and test version of the Reuters R8 collection and, partially, also on a reduced version of the 20 Newsgroups. In Figure 1 we may see the possible correlation between the relative hardness of the (i) train and (ii) test versions of the R8 collection with respect to the F-Measure obtained by using the MajorClust clustering algorithm. The smaller is the value of RH (x-axis) the higher is the obtained F-Measure (y-axis) and viceversa for both corpora. The relative hardness vs. $F$-measure correlation was calculated for all possible corpora variants of R8 (247).



(a) Train

(b) Test

**Fig. 1.** Evaluation of all R8 subcorpora (more than two categories per corpus)



(a) Train

(b) Test

**Fig. 2.** Evaluation of single pairs of the R8 categories

**Table 2.** The most related categories of the R8 collection

(a) Train

| RH value | Category | Category |
|---|---|---|
| 0.426 | trade | monex-fx |
| 0.399 | monex-fx | interest |
| 0.367 | trade | crude |
| 0.362 | monex-fx | crude |
| 0.352 | trade | interest |

(b) Test

| RH value | Category | Category |
|---|---|---|
| 0.419 | monex-fx | interest |
| 0.364 | trade | monex-fx |
| 0.332 | trade | interest |
| 0.317 | trade | crude |
| 0.311 | monex-fx | crude |

**Table 3.** The least related categories of the R8 collection

(a) Train

| RH value | Category | Category |
|---|---|---|
| 0.188 | interest | earn |
| 0.180 | acq | ship |
| 0.173 | ship | earn |
| 0.153 | grain | acq |
| 0.147 | grain | earn |

(b) Test

| RH value | Category | Category |
|---|---|---|
| 0.186 | interest | acq |
| 0.154 | ship | earn |
| 0.147 | acq | ship |
| 0.128 | grain | earn |
| 0.111 | grain | acq |

**Table 4.** The most related categories of the Mini20Newsgroups collection

| RH value | Category | Category |
|---|---|---|
| 0.3412 | talk politics guns | talk politics misc |
| 0.3170 | alt atheism | talk religion misc |
| 0.3092 | talk politics guns | talk religion misc |
| 0.3052 | talk politics misc | talk religion misc |
| 0.3041 | soc religion christian | talk religion misc |
| 0.2988 | sci crypt | talk politics guns |
| 0.2985 | soc religion christian | talk politics misc |
| 0.2958 | soc religion christian | talk politics guns |
| 0.2932 | talk politics mideast | talk politics misc |
| 0.2905 | sci electronics | sci space |
| 0.2868 | comp sys ibm pc hardware | comp sys mac hardware |

In order to easily visualise the correlation between RH and F-Measure, we have plotted the polynomial approximation of degree one.

In Figure 2 we may see the possible correlation between the relative hardness of each pair of categories of the R8 collection and the F-Measure obtained again by using the MajorClust clustering algorithm. The same conclusion is obtained: the smaller is the value of RH (x-axis) the higher is the obtained F-Measure (y-axis) and viceversa.

In order to fully appreciate the RH formula, the most and least related pairs of categories for the R8 dataset are presented in Tables 2 and 3, respectively. The RH value associated with each pair was calculated with the same formula presented in Section 3. Some preliminary experiments were carried out also with the Mini20Newsgroups dataset and the most and least related pairs of categories are shown in Tables 4 and 5, respectively.

**Table 5.** The least related categories of the Mini20Newsgroups collection

| RH value | Category | Category |
|---|---|---|
| 0.1814 | comp os mswindows misc | rec sport hockey |
| 0.1807 | misc forsale | talk politics misc |
| 0.1804 | misc forsale | talk religion misc |
| 0.1803 | comp sys ibm pc hardware | talk politics mideast |
| 0.1798 | comp os mswindows misc | talk religion misc |
| 0.1789 | alt atheism | comp os mswindows misc |
| 0.1767 | alt atheism | misc forsale |
| 0.1751 | misc forsale | soc religion christian |
| 0.1737 | comp os mswindows misc | soc religion christian |
| 0.1697 | misc forsale | talk politics mideast |
| 0.1670 | comp os mswindows misc | talk politics mideast |

## 6 Conclusions

We have observed that it is possible to introduce a measure to determine the relative hardness of clustering corpora based on the vocabulary overlapping. The obtained results show that there exists a correlation between the $F$-measure and the RH formula. With respect to the analysis carried out in [1], the introduced formula in our research work relies only on the vocabulary overlapping and it does not use any classifier. In fact, we use the MajorClust clustering algorithm only to evaluate the quality of the proposed formula by employing the $F$-measure. Therefore, the introduced RH formula may be used in an unsupervised manner in order to determine the relative hardness of clustering corpora.

## 7 Further Work

As future work, we need to investigate the correlation between the relative hardness and the F-Measure also on the Mini20Newsgroups dataset. Moreover, we are interested in evaluate both, the vocabulary overlapping and the term frequencies. This will allow us to further investigate whether the use of the tf-idf formula in the same context improves the current results or not. Besides, we would like to investigate the possible relationship the RH-Measure could have with cluster validity measures, such as the Density Expected Measure (DEM) which quantifies the similarity within clusters [**?**]. Moreover, we plan to determine the correlation between RH-Measure and the F-Measure through rank correlation coefficients such as Spearman's and Kendall's ones [4]. The final aim of this research work is to determine the level of hardness of a narrow-domain corpus, such as hep-ex [9], from a clustering task perspective.

## Acknowledgements

We would like to thank Dr. Mikhail Alexandrov from the Autonomous University of Barcelona for his invaluable feedback during the development of this research. This work has been partially supported by the MCyT TIN2006-15265-C06-04 research project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

## References

1. Debole, F., Sebastiani, F.: An analysis of the relative hardness of reuters-21578 subsets. Journal of the American Society for Information Science and Technology 56(6), 584–596 (2005)
2. Zaïane, O.R.: Principles of knowledge discovery in databases - Ch. 8: Data clustering (1999), online-textbook `http://www.cs.ualberta.ca/zaiane/courses/cmput690/slides/Chapter8/`
3. Meyer zu Eissen, S., Stein, B.: Analysis of clustering algorithms for web-based search. In: Karagiannis, D., Reimer, U. (eds.) PAKM 2002. LNCS (LNAI), vol. 2569, pp. 168–178. Springer, Heidelberg (2002)
4. Meyer zu Eissen, S.: On Information Need and Categorizing Search. Dissertation, University of Paderborn (2007)
5. Bezdek, J.C., Pal, N.R.: Cluster validation with generalized dunn's indices. In: 2nd International two-stream conference on ANNES, pp. 190–193 (1995)
6. Bezdek, J.C., Li, W.Q., Attikiouzel, Y., Windham, M.: Geometric approach to cluster validity for normal mixtures. Soft Computing 1(4), 166–179 (1997)
7. Stein, B., Nigemman, O.: On the nature of structure and its identification. In: Widmayer, P., Neyer, G., Eidenbenz, S. (eds.) WG 1999. LNCS, vol. 1665, pp. 122–134. Springer, Heidelberg (1999)
8. Kang, N.O., Gelbukh, A., Han, S.Y.: Ppchecker: Plagiarism pattern checker in document copy detection. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 661–667. Springer, Heidelberg (2006)
9. 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)