A fast dendrogram refinement approach for unsupervised expansion of hierarchies

Ricardo M. Marcacini, Everton A. Cherman, Jean Metz, and Solange O. Rezende

University of São Paulo (USP),
Mathematical and Computer Science Institute (ICMC)
São Carlos, SP, Brazil
{rmm,echerman,solange}@icmc.usp.br
jean@metzz.org

Abstract. Hierarchies are effective data models for organizing textual collections, particularly for automatic document classification into categories and subcategories. However, the majority of existing methods on hierarchical classification require human-labeled document set. Moreover, humans have good insight to manage the categories of higher levels of the hierarchy, *i.e.*, more general categories, while the management of more specific categories is a difficult and expensive task since it requires expert knowledge to identify appropriate categories and their respective documents. Thus, in this paper we introduce an approach to automatically expand new, and more specific categories from a reduced initial hierarchy, which contains only general categories. Our approach is based on text clustering methods, particularly performing refinements on dendrograms obtained by hierarchical clustering algorithms. The results of the experimental evaluation show that the proposed approach achieves better performance in the expansion of hierarchies, compared with a traditional technique. Moreover, our approach is computationally faster, allowing the identification of new categories in large text collections.

Keywords: clustering, expansion of hierarchies, refinement learning

1 Introduction

Hierarchies are effective data models for organizing textual collections, particularly for automatic document classification into categories and subcategories [12, 1]. Most existing studies on hierarchical classification have focused on investigation of methods to induce classifiers based on a previously available hierarchical structure. This hierarchical structure and its documents are usually labeled by human inspection, which requires a great human effort and is limited to a small set of documents [8]. Moreover, humans have good insight to manage the categories of higher levels of the hierarchy, *i.e.*, more general categories, while the management of more specific categories is a difficult and expensive task since it requires expert knowledge to identify appropriate categories and their respective documents.

The above scenario leads to an interesting research challenge as follows: from a reduced initial hierarchy that contains only general categories, the goal is to expand some categories into new subcategories in an unsupervised way. This approach is relevant for several applications, since it allows us to obtain more specific hierarchical models using the same human effort needed to obtain a more general hierarchical model. In other words, given a small sample of documents organized hierarchically by a human, then unsupervised machine learning methods, such clustering algorithms, are applied to identify new and more specific subcategories.

In this paper, we present an approach for unsupervised expansion of hierarchies based on hierarchical clustering. Hierarchical clustering algorithms organize a document collection into clusters and subclusters according to the similarity among documents. In general, hierarchical clustering obtains a binary tree called dendrogram that represents relationships between cluster and subclusters. We introduce a technique for dendrogram refinement, where inappropriate relationships between clusters and subclusters are identified and removed. The proposed approach was applied to automatically expand general categories of a previously existing reduced hierarchy on four text collections. The results obtained from the experimental evaluation show improvements on accuracy of expanded hierarchies when compared with a traditional technique. Moreover, our approach is computationally faster, allowing the identification of new categories in large text collections. Another experimental evaluation, involving large text collections, was submitted to the "Third Pascal Challenge on Large Scale Hierarchical Text classification (ECML/PKDD Discovery Challenge 2012)" obtaining good results in the unsupervised refinement learning task.

The rest of this paper is organized as follows. Section 2 presents the basic concepts of hierarchical text clustering used in this work. The proposed dendrogram refinement approach for unsupervised expansion of hierarchies is described in Section 3. Section 4 presents an experimental evaluation of our approach, describing the evaluation criteria, as well as analyzing and discussing the results. Finally, Section 5 presents the conclusions and directions for future work.

2 Background

In this work, an unsupervised machine learning process, based on hierarchical text clustering, is used to support the automatic expansion of class hierarchies. Therefore, it is necessary to choose a structured text representation model, a similarity measure among documents, and a clustering strategy.

We use the vector space model, which is one of the most common structures for text representation [1]. In this model, each document is represented by a vector of terms $x = (t_1, t_2, ..., t_m)$, where each term $i = 1..m$ has a t_i value associated to its relevance (weight) to the document. A document cluster $G = \{x_1, x_2, ..., x_n\}$ also has a representation in the vector space model, which is defined by a centroid $C_G = \frac{1}{|G|} \sum_{i=1}^{n} x_i$ (the mean vector of all documents belonging to cluster G).

The similarity between two documents x_i and x_j (or document clusters) represented in the vector space model is usually calculated using the cosine measure, which is defined as $cos(x_i, x_j) = \frac{x_i \times x_j}{||x_i|| \times ||x_j||}$ and results values in the interval $[0, 1]$. The cosine value between two documents is 1 whenever the documents are identical and 0 when they do not share any term (orthogonal vectors). In some cases, it is useful to adapt the cosine similarity measure to a dissimilarity measure by using the equation $dis(x_i, x_j) = 1 - cos(x_i, x_j)$.

The third aspect to choose is the hierarchical clustering strategy, which can be classified as agglomerative or divisive [11]. In the agglomerative hierarchical clustering, initially each document is a singleton cluster and, in each iteration, the closest pair of clusters is unified, until they form only one cluster. The divisive hierarchical clustering, on the other hand, starts with a unique cluster containing all documents, and is iteratively divided into smaller clusters until only singleton clusters remains. Experimental evaluations show that the algorithm UPGMA (agglomerative) and the Bisecting-kmeans (divisive) achieve the best results in textual data [15]. In both strategies, the results of hierarchical clustering are represented by a dendrogram.

The use of text clustering algorithms to expand categories into new subcategories is more common in automatic categorization of search results [3]. In this case, the results of a search engine are grouped together and presented to users through a list of topics (clustering results). Then, the user selects a topic (cluster) of interest, which is expanded into new subtopics. The refinement of hierarchical clustering has also been studied in the generation of taxonomies [4], which depends on the interaction with users to identify more comprehensive hierarchical structures. In this work, we propose an approach to expand a class hierarchy into more specific subcategories based on objective cluster validation measures [7]. Therefore, we assume that there are no users to identify which categories will be expanded, or even no information about the number of levels to expand.

3 An Approach for Fast Dendrogram Refinement

The method for expansion of hierarchies used in the proposed approach is illustrated in the Figure 1. First, a general category is selected to be expanded into subcategories (Figure 1(A)). Then, a hierarchical clustering algorithm is applied from the documents belonging to the category for the dendrogram extraction (Figure 1(B)). Finally, the dendrogram is refined to identify further subcategories and to complete the expansion of the hierarchy (Figure 1(C)).

Due to its usefulness to display de sequence of clusters and to identify similarities among clusters and subclusters, the dendogram is the most frequently used structure to represent hierarchical clustering algorithms results. However, the direct use of the dendrogram is not suitable for expansion of hierarchies due to the large number of clusters and subclusters obtained through such strategy. For example, to expand a hierarchy from a category with n documents, the resulting dendrogram contains at least $(n - 1)$ new subcategories. Therefore, the

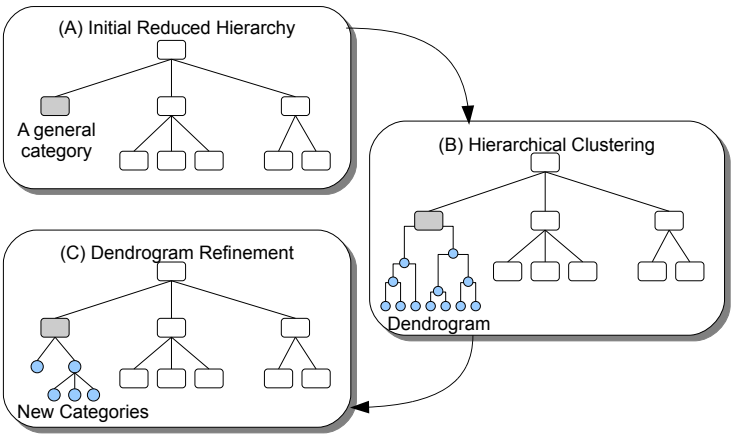

Fig. 1. The proposed approach for expansion of class hierarchies.

use of an approach to refine the structure of the dendrogram is crucial to expand hierarchies in the real world.

The proposed dendrogram refinement is based on clustering validation measures [2]. There are several criteria for clustering validation, but the most often used is related to the concept of cluster cohesion [13], which is defined by Equation 1, where G is a document cluster, C_G is a cluster vector and d_i is a document vector. Considering this criterion, the higher the cohesion value the better the cluster quality. Consequently, a cluster has good cohesion when the documents in the cluster are highly similar.

$$cohesion(G) = \frac{1}{n} \sum_{i=1}^{n} cos(d_i, C_G) \tag{1}$$

A limitation of the traditional cohesion criteria is that the hierarchical relationships are not considered in the cluster analysis. To avoid this problem, we propose the use of an alternative criterion called fusion, defined in Equation 2, where C_G^{PARENT} is the centroid of the parent cluster of the cluster G.

$$fusion(G) = dis(C_G, C_G^{PARENT}) \tag{2}$$

The basic idea of the fusion criteria for dendrogram refinement is described as follows:

a low fusion value indicates high similarity between the parent cluster and child cluster. This is an inappropriate relationship, since the child cluster G does no specialize the information already contained in the parent cluster (both describe very similar patterns). On the other hand, a high fusion value indicates that parent and child clusters are well separated (high dissimilarity) and both describe distinct patterns, which means that the cluster G is representative in the hierarchy and represent an appropriate specialization relationship.

After defining the quality criterion we can remove all inappropriate clusters from the dendogram, *i.e.*, the ones with the quality value that does not achieve a minimum threshold. When an inappropriate cluster is removed, its children clusters are promoted to the nearest parent cluster. After removing all the inappropriate clusters, a refined hierarchy is extracted from the dendrogram. This process is applied for each dendrogram obtained from the documents belonging to the general categories of the hierarchy.

The threshold value for the cluster quality criterion depends on the user's needs and on the characteristics of the textual collections. More details on the configuration of this parameter is discussed based on the results of the experimental evaluation.

4 Experimental Evaluation

We carried out an experimental evaluation to analyze the proposed approach for dendrogram refinement. We compared two criteria for the removal of inappropriate clusters from dendrograms: the "cluster cohesion" that is traditionally used in cluster validation, and the "fusion" that we have proposed as an alternative for dealing with hierarchical structures. The textual collections, as well as the algorithms for dendrogram refinement, are available online at `http://sites.labic.icmc.usp.br/torch/lshtc3`.

4.1 Textual Collections

We used a total of 4 benchmark textual collections from different sources. Table 1 presents a summary of these textual collections. The smallest collection contains 2,301 documents, while the largest collection contains 7,674 documents. All text documents were preprocessed by following the recommendations of [10]: (i) *stopwords* removal, such as pronouns, prepositions and articles; (ii) term *stemming* by Porter's algorithm, in which variations of a word are reduced to their radical; and (iii) removal of terms that occur in less than two documents.

Table 1. Summary of textual collections used in the experimental evaluation.

Dataset	Source	#Terms	#Documents	#Categories
ACM	Computer science papers	3,462	3,498	40
Hitech	San Jose Mercury (TREC)	2,289	2,301	6
LATimes	LA Times (TREC)	6,141	6,279	6
Re8	Reuters-21578	7,555	7,674	8

The documents in all textual collections are organized into categories of reference. Thus, it is possible to evaluate the performance of the dendrogram refinement approach by using objective measures, like the one addressed in the next section.

4.2 Evaluation Criteria

The F_{SCORE} index is a measure that uses the ideas of precision and recall from information retrieval to evaluate hierarchical clustering solutions. These measures are used as external validation criteria, because it uses prior knowledge to evaluate the clustering results. The rationale is to assess how the hierarchical clustering recovers the category information associated with each document. For this purpose, the following notation is used:

- H is a hierarchical clustering structure (dendrogram);
- L_r is a category of reference and its respective set of documents; and
- G_i is a cluster belonging to H and its respective set of documents.

Given a category L_r and a cluster G_i, we calculate the precision P and recall R according to Equation 3 and 4, respectively. Then, the harmonic mean F is obtained using the Equation 5, which is a balance between the precision and recall.

$$P(L_r, G_i) = \frac{|L_r \cap G_i|}{|G_i|} \quad (3) \qquad R(L_r, G_i) = \frac{|L_r \cap G_i|}{|L_i|} \quad (4)$$

$$F(L_r, G_i) = \frac{2 * P(L_r, G_i) * R(L_r, G_i)}{P(L_r, G_i) + R(L_r, G_i)} \quad (5)$$

The measure F selected for a particular category L_r is the highest value obtained for a cluster of the hierarchy H, according to Equation 6. Finally, the F_{SCORE} index of the hierarchical clustering with n documents and c categories, is the overall sum of F weighted by the number of documents (Equation 7). The higher the value of the F_{SCORE} index (which is in the range [0,1]), the better hierarchical clustering.

$$F(L_r) = \max_{G_i \in H} F(L_r, G_i) \quad (6)$$

$$F_{SCORE} = \sum_{r=1}^{c} \frac{|L_r|}{n} F(L_r) \quad (7)$$

We have adopted the F_{SCORE} index because it allows to evaluate our approach in an interesting way: the value of F_{SCORE} index is maximum for the original dendrogram, since it contains all cluster candidates for the expansion of the hierarchy. Thus, we consider that the dendrogram refinement is good if the F_{SCORE} index has little reduction after removing the inappropriate clusters.

4.3 Experiment Setup

We performed simulations where each textual collection represents a node of a hypothetical hierarchy to be expanded. We used the Bisecting K-means algorithm to perform the hierarchical clustering, therefore obtaining one dendrogram for each textual collection. Then, the proposed dendrogram refinement approach

was applied, and we compared the traditional cluster cohesion criteria with the proposed alternative fusion criteria.

We also performed several variations on the threshold value for the cluster quality criterion. In order to avoid scaling problems, the values are standardized using the well-known max-min standardization – the values range from 0 to 1.

Finally, the execution time of each run was recorded to compare the computational time efficiency of the cluster cohesion criteria with the alternative fusion criteria.

4.4 Results and Discussion

The experimental results are presented and discussed considering two aspects: (1) dendrogram refinement ratio, and (2) average execution time.

The dendrogram refinement ratio indicates how many clusters were removed from the original dendrogram. Figure 2 presents the dendrogram refinement ratio for each text collection. We compared the fusion criteria with the cohesion criteria through the F_{SCORE} index. One can observe that, the proposed fusion criteria is more robust regarding the dendrogram refinement ratio. For example, considering the dendrogram refinement ratio value of 0.6 from the ACM dataset – 60% of the clusters removed from dendrogram – the F_{SCORE} index value is 0.30 for the fusion criteria, while the F_{SCORE} has a value of 0.08 for the cohesion criteria. This means that the fusion criteria is more effective to identify inappropriate clusters. Moreover, this behavior is repeated in other textual collections.

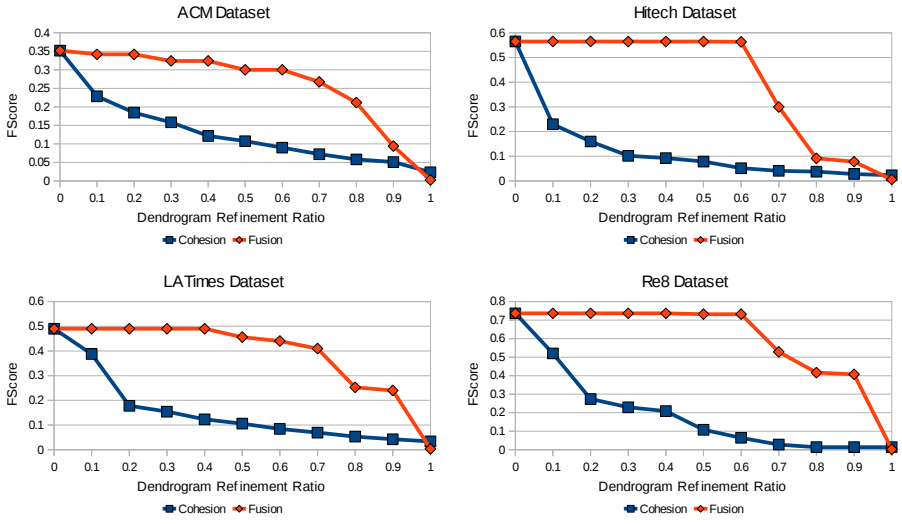

Fig. 2. Comparison of the cluster quality criteria for dendrogram refinement.

It is important to observe that when the dendrogram refinement ratio is high, then many clusters are removed from the dendrogram. Consequently, the remaining clusters may not be sufficient to represent all categories of reference, thereby significantly decreasing the F_{SCORE} index.

The dendrogram refinement ratio is obtained according to the cluster quality threshold value. Figures 3 and 4 present the relationship among the quality of the hierarchy, the dendrogram refinement ratio, and the threshold value. These results indicate that the fusion criteria is more aggressive than the cohesion criteria for the removal of inappropriate clusters without affecting the quality of the hierarchy.

A second aspect of evaluation is the average execution time shown in Table 2. It is noteworthy that the fusion criteria is faster than the traditional cohesion criteria, because the fusion criteria performs only a single calculation of similarity between two centroids, while the cohesion criteria calculates the similarity among all documents in a cluster and the cluster centroid.

Table 2. Comparison of the average execution time (miliseconds) between the cohesion and fusion criteria.

	Cohesion	**Fusion**
ACM	1762.10 ± 387.07	99.10 ± 9.94
Hitech	1427.45 ± 284.50	89.15 ± 8.09
LATimes	2983.75 ± 217.51	93.10 ± 7.69
Re8	2902.80 ± 143.56	97.35 ± 9.97

The experimental results were statistically compared by using the nonparametric Wilcoxon matched-pairs signed-ranks test with a 95% confidence interval. The analysis revealed that the fusion criterion has superior performance for the cohesion criteria with statistically significant difference in both F_{SCORE} values and average execution time. Thus, these results provide evidence that the dendrogram refinement approach with the fusion criteria has great potential to support the unsupervised expansion of hierarchies, especially in large textual collections.

4.5 Experiments with Large Scale Hierarchical Classification

We also applied the proposed dendrogram refinement approach to a large scale hierarchical classification task. The used textual collection and hierarchy was provided by Pascal LSHTC Challenge organizers (ECML / PKDD Discovery Challenge 2012). The goal is to expand – in an unsupervised manner – a hierarchy with approximately 12,000 nodes (categories), containing about 383,000 documents with 540,000 features (words).

We performed a hierarchical classification of approximately 103,000 new test examples (documents), which was also provided by Pascal LSHTC Challenge

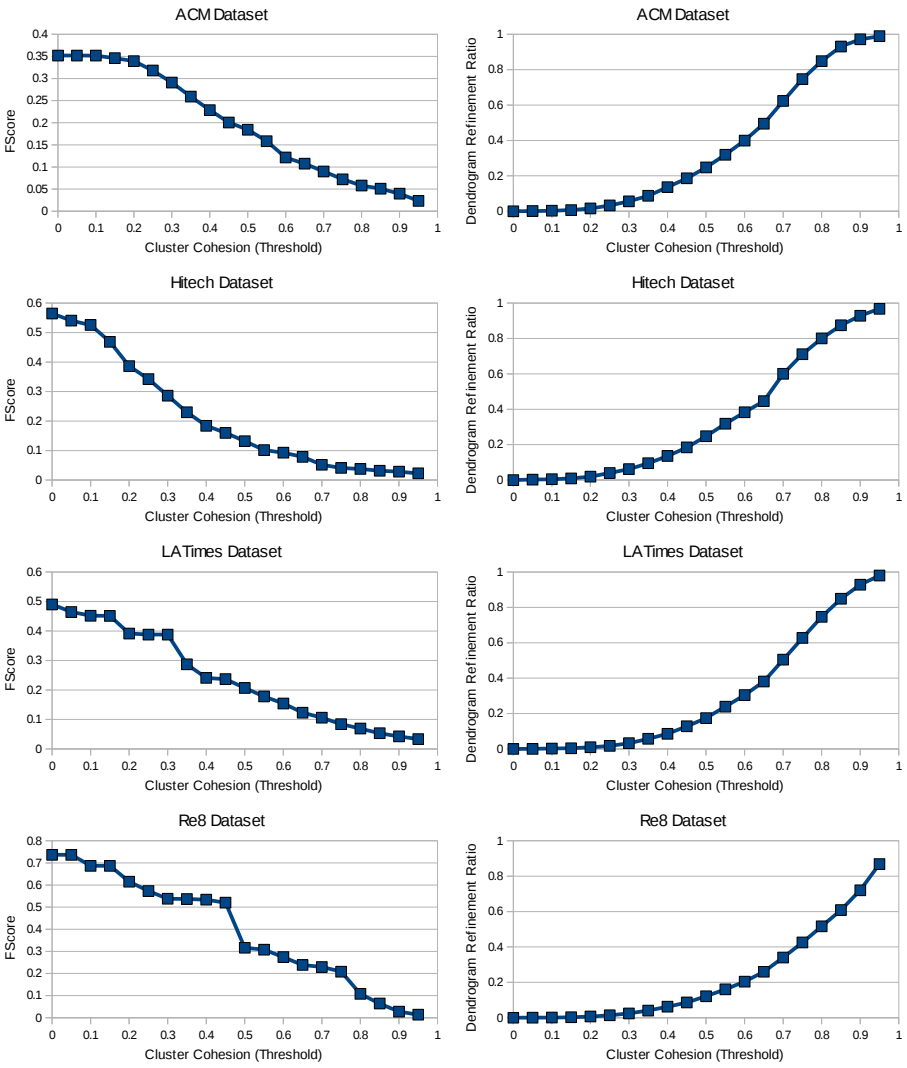

Fig. 3. Analysis of cluster quality threshold of the cohesion criteria.

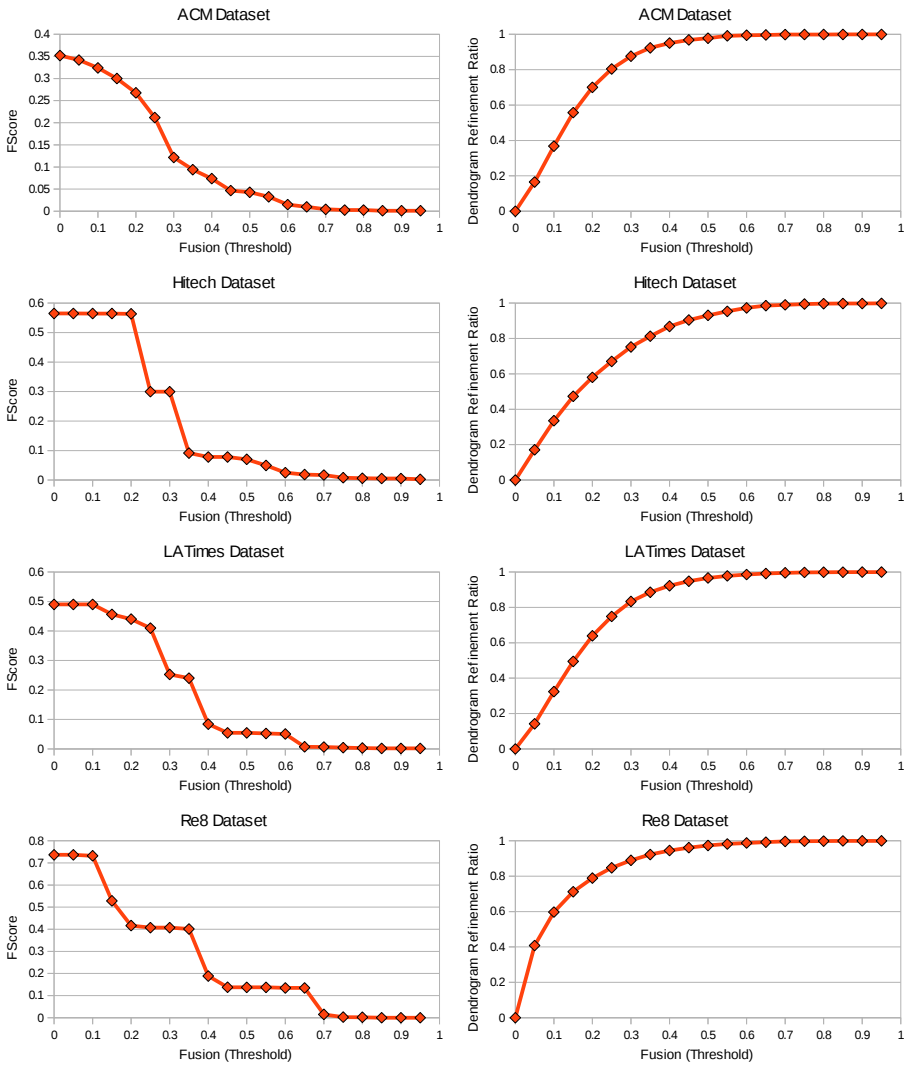

Fig. 4. Analysis of cluster quality threshold of the fusion criteria.

organizers. For this purpose, we used the hierarchical centroid-based classification algorithm [6, 9], because it has a low computational cost and uses the same document similarity concepts of the text clustering methods. Second, the leaf categories were selected for the expansion of the hierarchy. Therefore, we obtain dendrograms from the documents of each leaf category. Our dendrogram refinement approach was used to obtain the new subcategories of the expanded hierarchy.

After the unsupervised expansion of the hierarchy, we submit the results for the online evaluation system of the Pascal LSHTC Challenge and our approach obtained the following values: F = 0.36, Precision = 0.84, Recall = 0.28. It is important to note that these measures are based on ontology alignment [14].

5 Concluding Remarks

We presented a dendrogram refinement approach for unsupervised expansion of hierarchies. Two cluster quality criteria were evaluated to remove inappropriate clusters from a dendrogram: the cohesion criterion that is traditionally used for clustering validation, and the fusion criteria that was proposed in this paper as an alternative for the hierarchical scenario. The reported experimental results show that the proposed approach using the alternative fusion criteria yields to more robust dendrogram refinement, as well as providing a fast expansion of class hierarchies.

The unsupervised expansion of hierarchies is a recent challenge, with few studies addressing this issue. In this context, the proposed approach is simple and can be extended to many applications. For further study, we plan to incorporate information of the hierarchy and its corresponding training set to learn a more appropriate similarity measure [5]. We expect that a similarity measure learned from labeled data can improve the performance of the fusion criteria in identifying inappropriate clusters. Furthermore, we intend to perform a qualitative human evaluation of the quality of the expanded hierarchies.

6 Acknowledgements

The authors wish to thank FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) for financial support (process numbers 2010/20564-8, 2011/19850-9, 2010/15992-0, and 2011/21723-5).

References

1. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Mining Text Data, pp. 163–213. Springer-Verlag (2012)
2. Bouguessa, M., Wang, S., Sun, H.: An objective approach to cluster validation. Pattern Recognition Letters 27(13), 1419 – 1430 (2006)
3. Carpineto, C., Osiński, S., Romano, G., Weiss, D.: A survey of web clustering engines. ACM Computing Surveys (CSUR) 41(3), 17:1–17:38 (2009)

4. Chuang, S.L., Chien, L.F.: Taxonomy generation for text segments: A practical web-based approach. ACM Transactions on Information Systems (TOIS) 23(4), 363–396 (2005)
5. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: International conference on Machine Learning (ICML). pp. 209–216. ACM, New York, NY, USA (2007)
6. Han, E., Karypis, G.: Centroid-based document classification: Analysis and experimental results. Principles of Data Mining and Knowledge Discovery pp. 116–123 (2000)
7. Langfelder, P., Zhang, B., Horvath, S.: Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24(5), 719–720 (2008)
8. Li, X., Kuang, D., Ling, C.: Active learning for hierarchical text classification. In: Advances in Knowledge Discovery and Data Mining, vol. 7301, pp. 14–25 (2012)
9. Miao, Y., Qiu, X.: Hierarchical centroid-based classifier for large scale text classification. Large Scale Hierarchical Text classification (LSHTC) Pascal Challenge (2009)
10. Nogueira, B., Moura, M., Conrado, M., Rossi, R., Marcacini, R., Rezende, S.: Winning some of the document preprocessing challenges in a text mining process. In: IV Workshop on Algorithms and Data Mining Applications. pp. 10–18 (2008)
11. Rokach, L.: A survey of clustering algorithms. In: Data Mining and Knowledge Discovery, pp. 269–298. Springer, 2nd edn. (2010)
12. Sun, A., Lim, E.P.: Hierarchical text classification and evaluation. In: IEEE International Conference on Data Mining (ICDM). pp. 521 –528 (2001)
13. Vendramin, L., Campello, R.J.G.B., Hruschka, E.R.: Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining 3(4), 209–235 (2010)
14. Zavitsanos, E., Paliouras, G., Vouros, G.: Gold standard evaluation of ontology learning methods through ontology transformation and alignment. IEEE Transactions on Knowledge and Data Engineering pp. 1635 – 1648 (2010)
15. Zhao, Y., Karypis, G., Fayyad, U.: Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery 10(2), 141–168 (2005)