Easily Labelling Hierarchical Document Clusters

Maria Fernanda Moura[1,2], **Ricardo Marcondes Marcacini**[2], **Solange Oliveira Rezende**[2]

[1]Embrapa Informática Agropecuária - Campinas - SP - Brazil

[2]Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo
São Carlos - SP - Brazil

fernanda@cnptia.embrapa.br, marcacini@grad.icmc.usp.br, solange@icmc.usp.br

Abstract. *One of the problems of automatic models that generate topic taxonomies is the process of creating the most significant term list that discriminates each document group. In this paper, a new method to label document hierarchical clusters is proposed, which is completely independent from the clustering method. This method automatically decides the number of the words in each label list, avoids word repetitions in a tree branch and provides a kind of cutting for the cluster tree. The obtained results were tested as search queries in a retrieval process and showed a very good performance. Additionally, the use of the method was experimented by some specialists in the text collection domain, trying to evaluate their understanding and expectations over the results.*

1. Introduction

Labelling clusters is a common problem in text mining and information retrieval. Generally, the methods find a list of discriminative words, that are used to facilitate the information retrieval or the interpretation of the groups. The results could be used as the first step to aid in the construction of a topic taxonomy, since the documents are from a specific domain and a domain specialist is involved in the task. The topic taxonomy is helpful in organizing documents, for example, aiding a digital library or a portal building up.

Although there are very good methods dependent on a specific cluster algorithm, we can treat the hierarchical cluster labelling problem as a supervised or semi supervised attribute selection [Weiss et al. 2005]. There are many proposals, that follow the Glover´s ideas based on the observed frequencies for each term t in the collection $p(t)$ and in each group $p(t/g)$ [Glover et al. 2002]. The assumed hypothesis is that if $p(t/g)$ is very common and $p(t)$ is rare then the term is a good discriminator for the *g* class, or even $p(t/g)$ and $p(t)$ are common so the term discriminates the *parent class of g* and, finally, if $p(t/g)$ is very common and $p(t)$ is relatively rare in the collection so the term is a better discriminator for the *child class of g*; the very common and rare thresholds are experimentally determined. A modification was proposed for this method, where there is a compromise between a simple label and a label list, establishing a *descriptive score* pondered by *tf-idf* [Treeratpituk and Callan 2006]. Although the results are good, the problem of experimentally determining the convergence criteria was spread to the new *cutoffs* needed for the *descriptive score*.

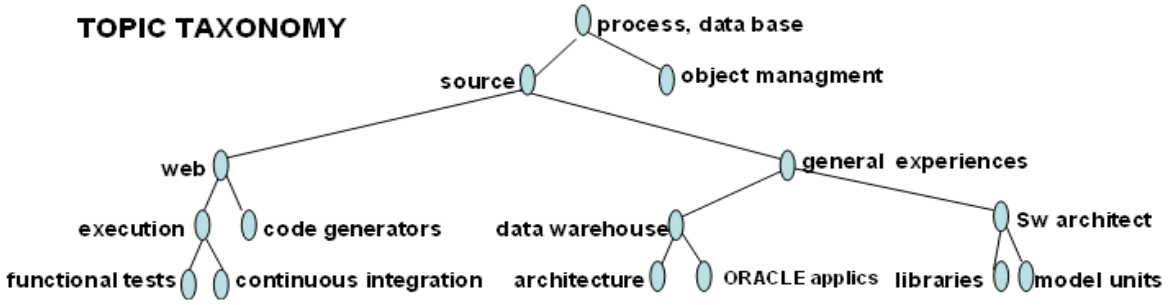

Figure 1. Topic taxonomy inferred to some papers of informatics applications

An older method that works over a given multinomial term distribution in a hierarchical grouping was developed by Popescul and Ungar [2000]. This proposal uses an attribute selection criteria testing each term dependence on the child nodes; if the independence hypothesis is accepted the term is related only to the parent node not to the children, else it belongs only to the children list; according to Glover´s assumptions. The advantage over Glover´s method is that this method does not need to train a threshold. In this work, we proposed a new method inspired by the Popescul and Ungar proposal. The proposed method is always able to make a decision about any term and generates a smaller label list for each cluster. It also avoids term repetitions along the hierarchy and provides an automatical cutting criteria to the cluster tree. Moreover, our method is cluster algorithm, domain and language independent.

In a previous study of the proposed cluster labelling method variations, the hierarchical grouping was obtained by some bottom up agglomerative hierarchical clustering algorithm and the labelling methods worked over the generated binary tree; these descriptions are found in [Moura and Rezende 2007]. The algorithm presented in this paper was expanded to any kind of hierarchy (not only a binary tree), choosing the needed decision estimates according to the children number of nodes of each tree node. All descriptions to elucidate the algorithm and its contributions are found in the methodology section. The preliminary results are very good. They were tested against an information retrieval process and submitted to a subjective analysis by some domain specialists, detailed in the experiment section. Despite the encouraging results, the method demands some improvements and future work to make the result interpretation easier, as discussed in the final considerations section.

2. Methodology

In this work, term can be either a word or a stemmized word, considered alone or in a phrase combination. The goal is to distribute the terms along the hierarchy, avoiding unnecessary repetitions in the same branches, keeping the most generic terms in the high levels and the most specific terms in the low levels. In the Fig. 1 we see an inference over some labels for some papers about agricultural informatics; in which the cluster labels were obtained with the method proposed here. Since some documents are in the same cluster, they are supposed to cover the same topic. Following the Fig. 1 the topic "source" probably refers to source code, that was divided into "web" code and "general experiences" in software production.

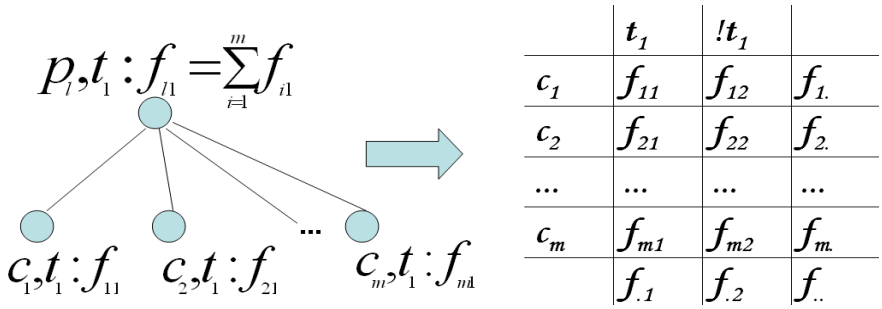

Figure 2. Term and its frequency in the parents and children nodes

2.1. General Idea and Definitions

Each hierarchical node corresponds to a list of terms presented in its documents, and for each term the cumulative absolute frequency is calculated. Considering each children group of a fixed node, the hypotheses of independence (or dissociation) are tested for each term in each group; considering the parent group as the current and the children as the tested groups. For example, in Fig. 2, one can observe a term t_1 which is presented in the parent node p_l and in its children set $[c_1, c_2, ..., c_m]$, with its respective frequency in each node $f_{1i},\ i = 1, .., m$. In order to decide if the term discriminates only the parent node or only one of the children or the children set, an independence test is applied over the term distribution in the children. To carry out the test, for each term in each node, the set of children is divided into two classes, according to the presence or the absence of the term in the child class, resulting in a contingency table as also illustrated in Fig. 2 for the fixed term t_1. Each cell of the contingency table corresponds to the following definitions, used along this work, considering $i = 1, ..., m$:

- f_{i1}: absolute cumulative frequency of the term t_1 in the i^{th} child;
- f_{i2}: absolute cumulative frequency of the other terms, $f_{i.} - f_{i1}$;
- $f_{i.}$: absolute cumulative frequency of all terms in the i^{th} child;
- $f_{.1}$: total of the absolute cumulative frequency of the t_1 term in the parent node;
- $f_{.2}$: total of the absolute cumulative frequency of the other terms in the parent node;
- $f_{..}$: total of the absolute cumulative frequency of all terms in the $parent$ node.

Under the hypothesis of independence, that is, the fixed term t_1 does not depend on the children, each cell f_{ij} is supposed to depend exclusively on the marginal frequencies; that is, $e_{ij} = f_{i.} * f_{.j} * f_{..}$ is the expected value for each f_{ij} cell. If the tested hypothesis is true, that is, the term (in this case t_1) does not depend on the children nodes, the term depends only on the parent node; else, the term depends on the children nodes.

Following these ideas, the original method proposed by Popescul and Ungar [2000] carries out a test for each term in the hierarchy from the root to the leaves of the cluster tree. If the term depends only on the parent node, it is removed from each child term lists and remains in the parent node term list; else the term is removed from the parent list and remains in the children lists. In the end of the process, the terms which remain in the node lists are the selected labels for those nodes. In order to test the hypothesis, the original method used the chi-square statistical distribution [Popescul and Ungar 2000].

There are some problems with this approach, because the constraints to apply the chi-square test involves the absence of low frequencies in each cell of the contingency table, which is not always true for term distribution in clusters in a text mining proccess.

In the original method, Popescul and Ungar used the constraint of $5 \geq e_{ij} \geq f_{ij}$ as widely indicated in the statistical literature. So, if the contingency table for a fixed term has some frequency or expected frequency less than 5, the method is not able to make a decision, consequentely the tested term remains in all term lists along the hierarchy from the actual node point. This restriction can be relaxed to $1 \geq e_{ij} \geq f_{ij}$ when the total term frequencies are very big, but it already depends on a chi-square distribution, that can not be guaranteed under adverse conditions of the term frequency distribution (for details see [Bishop et al. 1984]). To improve the method, using its best insights, it was necessary to find a good estimator to be used in the tests and treat the extreme conditions, as the $f_{ij} \approx 0$.

The first improvement was to change the used estimators, according to the constraints and the number of children in each actual node, looking for estimators that do not depend on a specific probability distribution. In $2x2$ contingency tables, when the current node has only two children, the chosen estimator was Yule Q. To test the hypothesis of association using the Yule Q, the cross-product ratio $\alpha = (f_{11} * f_{22})/(f_{12} * f_{21})$ has to be calculated, and then the Q estimate[1] (for details see [Bishop et al. 1984]):

$$\widehat{Q} = \frac{\widehat{\alpha} - 1}{\widehat{\alpha} + 1},\ with\ \widehat{\sigma}_Q = \frac{1}{2} * (1 - \widehat{Q}^2) * \sqrt{\frac{1}{f_{11}} + \frac{1}{f_{12}} + \frac{1}{f_{21}} + \frac{1}{f_{22}}} \tag{1}$$

$$\widehat{Q} \approx N(\widehat{Q}, \widehat{\sigma}_Q) \Rightarrow \widehat{Q} \pm 2 * \widehat{\sigma}_Q \tag{2}$$

The maximum value of the function is reached when $\widehat{\alpha} = 1$ and $\widehat{Q} = 0$ and, $\widehat{Q} = 1$ or $\widehat{Q} = -1$ occurs when some $f_{ij} = 0$; so, if the value $0 \in \left[\widehat{Q} - 2 * \widehat{\sigma}, \widehat{Q} + 2 * \widehat{\sigma}\right]$ then the independence hypothesis, or dissociation hypothesis, is true.

To expand the algorithm to $mx2$ contingency tables, that is, the current node can have any number ($m \geq 3$) of children in each child, the U^2 estimator is used:

$$U^2 = \frac{(m-1) * BSS}{TSS},\ with\ BSS = TSS - WSS \tag{3}$$

$$TSS = (m/2) - (1/2m) \sum_i f_{i.}^2\ \ and\ \ WSS = (m/2) - (1/2) \sum_j \frac{1}{f_{.j}} \sum_i f_{ij}^2 \tag{4}$$

The TSS is interpreted as the total variance in the table, or the total dispersion among the values. The WSS is the children variance within the class; a positive class corresponds to the presence of the term in the child node. The BSS is the children variance among the classes. So, U^2 is an estimator for the reduction in the proportion of explained variance of data, that is, the term frequency distribution variance, and is assymptotically approximated by a chi-square distribution with $(m-1)$ degrees of freedom in this work (see [Bishop et al. 1984] for details) and does not depend on the probability distribution of the term frequencies.

The second improvement considers the extreme conditions when f_{ij} is approximately zero. The number of children in the actual parent node considered for each term depends on the term frequency in each child. If the j^{th} term is presented in a i^{th} child of the actual parent node, that is, its $f_{ij} \geq 1$, then the i^{th} child is considered in the test. So, the children which have $f_{ij} \approx 0$ are considered as completely dissociated from

[1]Every estimate is noted with a hat.

the j^{th} term. Additionally, if the parent node has only one child with the term occurence, the independence test is not applied, the term is considered completely associated to that unique child (for details see [Bishop et al. 1984]).

The third improvement is a cutting over the cluster tree, which is a direct consequence of the first and second improvements. The improved method is always able to make a decision about a specific term and consequentely avoids term repetitions along the hierarchy. In this way, sometimes, the method does not find even one term to discriminate a node, that means the node has an empty discriminative term list. Experimentally, we noticed that this occurred because in the collection there was not any term to discriminate the specific group, or when the formed cluster had not a real meaning. In these cases empty term lists are produced and treated as an automatical cutting for the cluster tree. The cutting follows the idea that generic terms are in the parents and they refer to children, so the children of the empty list node is just moved to the children set of the gran parent node.

2.2. Evaluation Method

The evaluation of the proposed method against the original one is available after its implementation as a prototype (developed in C). The prototype receives a hierarchical description of the cluster results for a document collection and is able to create the label list sets of each hierarchical document group for the two different methods. The Popescul & Ungar [2000] method was implemented as proposed, using the chi-square estimator, restricting the p-value to 0.05 in each hierarchical level and with the restriction $5 \geq e_{ij} \geq f_{ij}$, to apply the test. The new proposed method was implemented as explained in the methodology section. It has to be noted that two sets of cluster labels are obtained for each cluster hierarchy over a text collection: one generated from the original method and one generated from the proposed method.

Since we have the results of the two methods, a subjective evaluation is applied by the domain specialists over a hierarchical visualizaton of the results. The evaluators are asked to set a grade for the label list set of each group, for each method. An even number of grades, from 1 to 4 was chosen; in this way it is possible to avoid the mean grade when the evaluator is in doubt. To compare the grades we carried out a statistical mean comparision based on t student estimator; the goal is to verify how much the effects of the different methods influence the grade mean estimate.

To reach an objective evaluation, the measures of precision and recall were obtained from a simple retrieval process. The retrieval process was implemented (as another prototype) over the attribute-value matrices used in the document clustering process. The attribute-value matrices are composed by the documents in the lines and the terms in the columns, having the absolute frequencies as the values for each attribute in each document. The search queries correspond to each label list generated for each method, considering the "and" operator among the terms. In order to decide if a document had been retrieved or not, the presence of each term of the list in the document is necessary, that is, $f_{ij} > 0$ for the j^{th} term in the i^{th} document. After the retrieval process in each node, the following values has to be calculated:

- tp: the number of documents retrieved with the query that really belong to the cluster;
- fp: the number of documents retrieved with the query that do not belong to the cluster;
- t_r: the total number of retrieved documents with the specified query;

- fn: the number of documents not retrieved with the query that belong to the cluster; and,
- t_c : the number of documents in the cluster;

The precision and recall measures are respectively defined as: $p = tp/t_r$ and $r = tp/t_c$, in a range from 0 to 1. To understand the distribution and the balance between these measures, their harmonic mean is calculated as $F_{score} = 2 * p * r/(p + r)$. The ideal value of the F_{score} is equal to one, because it had to have $p = 1$ and $r = 1$; but generally it is sufficient to have a harmonic behavior of F_{score} along its graphic.

3. Experiments and Results

First the evaluators were chosen among specialists in the domain text collection. So, a small subjectively significant sample of documents for each domain were established. The samples are small, because analysing subjectively and in details an extensive automatic generated taxonomy is not a trivial task and can result in a low quality evaluation. The first text collection was randomly chosen among scientific publications in Portuguese about Artificial Intelligence in a total of 47 complete articles. The second text collection was the complete set of computational linguistic from 2005 to 2007 of the TIL event (TIL - "Tecnologia da Informação e da Linguagem Humana") composed by 51 complete articles also in Portuguese.

In the preprocessing step, the same *stopwords* lists and the same stemming process were applied to each text collections separately, using the PreText tool [Matsubara et al. 2003]. Onegram representations of the stemmized words were created and their frequency were counted in each text. The filtering process was carried out based on Luhn cutoffs, observing the stem frequency graph, only to the stemmized words presented in at least two documents. The hierarchy was obtained from the attribute-value matrix, using the MatLab environment. The dissimilarity metric based on cosine and the *average linkage* algorithm were used in the bottom up clustering process. Finally, the labelling algorithms were performed over the hierarchies and the results were shown in a visualization process.

The results for a branch of the hierarchy of the artificial intelligence document collection by the original method of Popescul & Ungar and the new proposed method are shown in Fig. 3. In the results obtained by the original method, some terms are repeated along the children nodes, for example *"document, sentenc, ontolog, reticul..."* (document, sentence, ontology, lattice...); while in the results of the new proposal, the terms are not repeated and are more discriminative. One good example is to compare the results of both methods in the last children of the hierarchy, where *"document, ontolo, domin..."* (document, ontology and domain) correspond to *"sinonimia, ..."* and where *"text, sentence, reticul, extract..."* (text, sentence, lattice, extraction,...) correspond to *"summariz, estrutur, corresponden, ..."* (summarization, structure, correspondent,...). This example is an evidence that specifically for the proposed method, the most generic terms were left in the high nodes and the most specific are really in their corresponding nodes.

The subjective evaluations were carried out over each method result for each text collection. To compare the methods, the evaluators were divided into pairs. Each pair of evaluators set a grade to each label after observing the both results, so the obtained number of grades depends on the number of evaluators and nodes in each hierarchy. In Table 1 there are the final grade means ($\overline{g}$) and their standard errors (se) for each method,

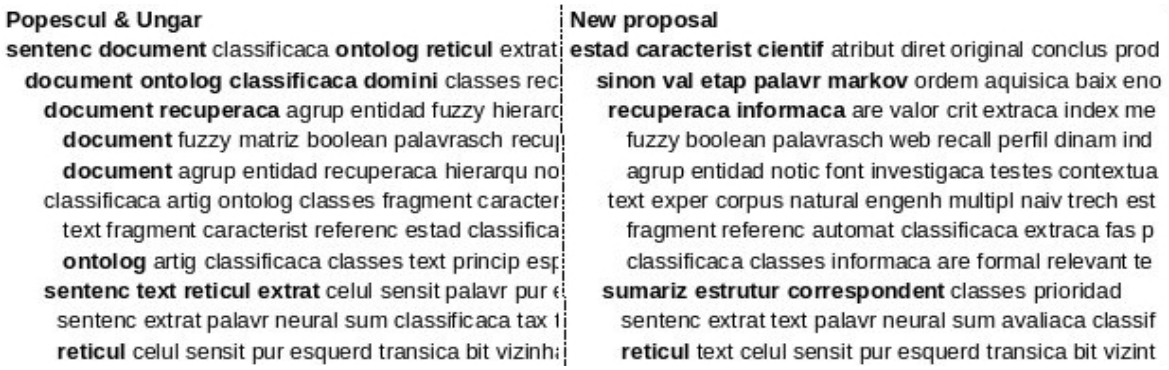

Figure 3. Comparing the results of the tow methods for the Artificial intelligence hierarchy.

$ArtificialIntelligence$			TIL		
$74\ grades$			$98\ grades$		
m	$\overline{g}$	se	m	$\overline{g}$	se
$P\&U$	2.7703	0.7860	$P\&U$	2.7449	0.7365
New	2.5811	0.9364	New	2.6020	0.6996

Table 1. Grade means ($\overline{g}$) and their standard errors (se) for each the method (m).

that were compared through a two tailed t student test. The calculated p-values were 0.0819 and 0.0713 for artifical intelligence and the TIL text collections respectively. In this way, we can conclude that the specialists did not find a difference in interpreting the label list meaning between both methods at a 5% significance level.

Some of the specialists suggested that the words repetition presented in the original method would have a broader interpretation and probably would be the best choice in a retrieval process. Apparently it seems to be a good choice, but it is good to discriminate clusters that does not have intersections, as for a k-means clusters results. If the clusters were completely independent, the results for a search query like "document or text or sentence" could be able to retrieve the goal documents. But in a hierarchy, the terms must be more specific, or all the documents in the collection will be retrieved, providing a $recall \approx 1$ and a very low precision. To test this hypothesis about retrieval and compare the results, the F_{score} points were calculated and plotted for both document collections, as shown in Fig. 4. For those F_{score} points, the precison and recall metrics were obtained from search queries, which considered only the ten first terms in each label list; because the domain specialists consider only the first terms as important.

Some of the F_{score} points could not be defined, because they resulted in a division by 0, when the recall and the precision tend to zero. In the graphics we can observe that the label lists produced by the new method provide better search queries than the lists produced by the original method. The method proposed here was responsible for 30 (65%) of the F_{score} points in the artificial intelligence F_{score} plot against the 5 (11%) points of the original method; and for the TIL collection, the proposed method had 37 (74%) of the F_{score} points against the 12 (24%) of the original method. Observing the graphics, the method proposed here finds the most discriminative lists of terms in each cluster, because it can answer the search queries in a bigger number of times and has

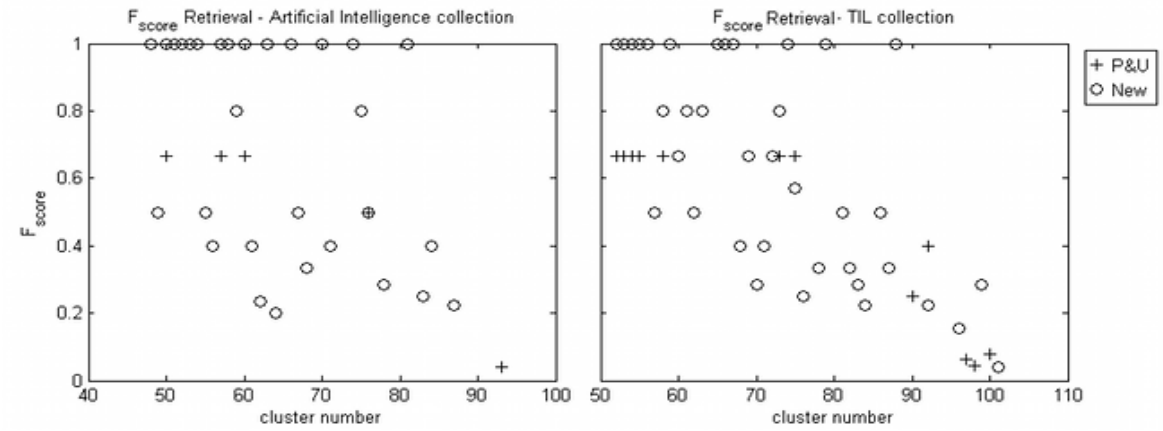

Figure 4. F_{score} for the retrieval process of the two collections.

better precisions and harmonic values of F_{score} points.

In conclusion the proposed method reached the best results, because it was comparable to the original in the subjective evaluation and it gave better results in the objective evaluation; additionally, it provided label lists with no intersections at each hierarchy level, that is, for all terms it had made a decision.

4. Final Considerations

In this work an automatic hierarchical cluster labelling method is proposed, aiming to adapt it to a topic identification process. The proposal is focused on the problem of avoiding term repetitions in the discriminative term sets along the hierarchy and on reducing the generated hierarchy. Not only the method reached those goals, but also the method does not depend on any threshold training or in a specific cluster algorithm, so it can be directly applied over any multinomial term distribution.

Besides the good evaluations of the domain specialists and the great results of the F_{score} for the proposed method, there is some future work to be done. The domain specialists complained about the absence of collocations in the term set lists; which can be done by integrating the use of n-gram words to the attributes. Probably, the use of n-gram words will help in the interpretation of the term lists, because it will add some semantic information to the bag of word approach. However, the generation of the term lists is only the first step to construct a topic taxonomy and to organize the document collection. In order to effectively help the topic taxonomy construction, the tool must allow the specialist´s intervention in constructing the branches and label sets, guiding him in the changes with the estimates obtained in each process step.

References

Bishop, Y., Fienberg, S. E., and Holland, P. H. (1984). *Discrete Multivariate Analisys*. MIT Press.

Glover, E., Pennock, D., Lawrence, S., and Krovetz, R. (2002). Inferring hierarchical descriptions. In *Conference on Information and Knowledge Management - CIKM*, pages 507–514.

Matsubara, E. T., Martins, C. A., and Monard, M. C. (2003). Pre-text: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Technical Report 209, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos.

Moura, M. F. and Rezende, S. O. (2007). Choosing a hierarchical cluster labelling method for a specific domain document collection. In *EPIA- Encontro Portugues de Inteligência Artificial, 2007, Guimarães, Portugual. New Trends in Artificial Intelligence. Lisboa, Portugual: APPIA - Associação Portuguesa para Inteligência Artificial*, pages 812–823.

Popescul, A. and Ungar, L. (2000). Automatic labeling of document clusters, unpublished manuscript (2000).

Treeratpituk, P. and Callan, J. (2006). Automatically labeling hierarchical clusters. In *Proceedings of the 7th Annual International Conference on Digital Government Research,San Diego, California, USA, May 21-24*, pages 167–176.

Weiss, S. M., Indurkhya, N., Zhang, T., and Damerau, F. J. (2005). *Text Mining - Predictive Methods for Analyzing Unstructured Information*. Springer Science+Business Media, Inc. ISBN 0-387-95433-3.