Category labelling for automatic classification scheme generation

Rodrigo Sánchez Jiménez
Dpto. de Biblioteconomía y Documentación UCM.
Facultad de Ciencias de la Información
Avda. Complutense, s/n. (Moncloa)
28040 Madrid, SPAIN

rsanchezj@ccinf.ucm.es

Abstract: This paper proposes a research line for developing new ways of automatically characterizing groups of documents, being them either clusters or categories. This research line is based upon the works of many other researchers and tries to summarize the most problematic issues of category labelling in order to devise possible solutions. Various lines of action are described, as well as future research lines and developments.

Keywords: Automatic classification scheme generation, document clustering, cluster labelling, text categorization, category labelling

1. MOTIVATION OF RESEARCH.

Classification schemes such as UDC or DEWEY have been successfully used for years in libraries around the world as information access tools. They provide us with a set of relations between concepts that help us to describe and retrieve documents. All purpose, fully fledged classification schemes as the ones mentioned above are extremely expensive to create and maintain, and suffer from a major drawback, lack of adaptation to collections. This last extent is derived from the fact that they have a universal scope and changes are slow, precisely because of the huge structure involved.

In many practical situations, taxonomies (purely hierarchical classification schemes), “folksonomies” (a set of flat, non-explicitly related and uncontrolled categories) or hierarchical directories have been used as a more tractable way to tackle document organization problems. This is the case of the ACM Computing Classification System, the del.icio.us folksonomy and the DMOZ open directory.

Classification schemes and related conceptual structures are the basis of a wide spread, user friendly way of exploring a collection structure and its contents, browsing. The main problem with browsing mechanisms is that creating them is a time consuming task, even in their less complex forms. This justifies the need of tools that enable the automation of the process. Our main goal is to provide a fast and unsupervised method for constructing collection tailored classification schemes.

Apart from generating browsable access to collections, the process of creating classification schemes from the basis of collection documents might have an interesting side-effect for automatic document categorization problems. Categories obtained in this fashion could also constitute an excellent source of training examples for categorizing new documents.

Automatic text categorization is strongly based upon machine learning algorithms, which provide the core for automatic categorization systems. This machine learning algorithms, are usually divided into two groups, profile based and example based algorithms [16]. Both of them need trained examples in order to correctly accomplish their goals. Trained examples should ascend to a respectable number of total documents of the collection to provide a non overfitting base for machine learning. The crux of the problem is thus an adequate supply of human designed categories and examples which implies quite a lot of human work for setting up classifiers. Both categories and training examples could be created during the classification scheme generation process.

The problem of automatically creating classification schemes or taxonomies has already been addressed by other researchers. In this paper we will focus on exploring the possibilities of improving the automatic description of groups of documents from a thematic point of view. Categories can be extracted and structured in a hierarchical fashion using state of the art methods and algorithms, identifying and labelling the main theme (or subject) of categories, is still a problematic point which has not been satisfactorily solved. Accurate and user friendly labelling should make more usable many techniques relying on clustering or automatic categorization methods.

2. BACKGROUND AND PREVIOUS WORK

In order to devise a method of creating both the classification scheme and document assignments to these categories we divided our work in three stages: category creation, category structuration and category naming. They are all based upon the idea that we can find a number of common characteristics in a given set of documents that distinguishes it from other sets of documents, much like in van Rijsbergen's cluster hypothesis [19]. These common characteristics are derived from underlying concepts, topics related to documents. We can assimilate the idea of finding groups of documents with the same thematic characteristics to that of extracting the main categories of a collection.

Category creation, or category extraction, issues the problem of finding groups of related documents according to a thematic criterion. This was accomplished by using document clustering techniques. The use of clustering algorithms is wide spread for many purposes, including that of unsupervised categorization. Non hierarchical algorithms (Expectation Maximization [17] and K-Means [20]) were problematic because of the fact that they needed certain information as the number of seeds and some knowledge on the collection structure [17]. Hierarchical algorithms as bisecting K-means [18] or HAC with single and complete link [20] were much more robust because they needed no initial information [17, 20]. Categories, or document clusters, are found at this stage of process, but have yet to be interpreted if they are to be of any use. Document clustering techniques are complex (many of them showing quadratic complexity), but organizing a collection is an offline, periodic routine.

The category structuration stage (or hierarchy interpretation stage) was derived from the fact that hierarchical algorithms create a dendrogram structure as a result of their work. That structure provides a hierarchy of nested links of documents and clusters reaching a number of hierarchical levels that was equal to $n-1$, being n the number of documents in the collection. Dendrogram interpretation (or folding) has been the subject of research of multiple papers [1, 3, 4]. Nevertheless, the question of how to fold the hierarchical structure and how many folds were optimal is still unsolved.

We devised a method for hierarchical clustering (the triple link algorithm) that needs no user interaction as to decide upon the number of folds or cuts to be made to the structure. It uses comparisons between documents to create a directed graph from which clusters are extracted. The process is repeated iteratively to create various hierarchical levels through a subsumption mechanism. This is still to be tested in depth, and it might not provide us with an optimal number of folds. On the other hand, any of the hierarchical approaches renders a group of thematically cohesive clusters with a good degree of success for a reasonable number of hierarchical levels.

Anyhow, category naming is still the biggest problem to be solved if any of these methods is to become useful for a final user. Once the categories are extracted, derived from groups of documents, and structure is given to them, the problem is as simple as not having an adequate way of naming them. Documents are presumed to be about the same topic, but we are unable to describe this topic (to name it) so as to bring forward a clear idea of the contents of each group of documents relating a same central topic. This is where cluster labelling methods come to be of great interest.

3. RELATED WORK

There are some very interesting papers on cluster labelling, although we think that there is a lot of research to be done in this field. We found three main lines of research, and will organize related work around them. The most basic approaches deal with the representation of categories by means of a series of terms. The second approach is derived from the first and attempts to obtain human-readable labels by means of submitting them to a post-process of user edition. The third and more elaborated way is related with the automatic extraction of a single label to represent the category with no user intervention.

3.1 Representing categories by a set of relevant terms

A good number of researchers just provide the final user with a group of highly relevant terms to describe the topics of the cluster [5, 2, 3, 5, 7]. In these cases the final user is presented with a list of terms that form the label. These labels are not usually "user-friendly", in the sense that they are related to a lot of small and in some cases unrelated topics. We can show here a sample from the LabelSOM system [8]:

Charl common market bomb greater missil france effort therefor interest polari weight stag europ gaul	T072 T071 T085 T086 T045
---	--

TABLE 1: A sample label from LabelSOM. Terms on the left are stemmed.

The main issue with these labels is that they are unable to provide us with a synthetic view of the contents. A correct understanding of the topics underlying this category rests upon previous knowledge of the collection, which is not supposed to be the case. There are probably best suited samples, but this one is explanatory. Some of the words are just stop words, and the set of terms would probably improve if key-phrases were used too, so this might still prove of some help in describing categories, although its far from constituting an ideal representation.

3.2 Representing categories by user-edited labels

Other researchers [9, 10] propose a method in which human editors are presented with a list of terms or candidate labels in order to provide the base for devising a label covering the central aspects of every category. If we give an expert in the dominion of the collection an adequate description of the categories, based upon the most relevant terms that characterize them, results can be quite good. This kind of manual post-edition gives us category labels that actually reflect the contents of the category and are understandable by the final user.

Manual editing is assisted by the system, which provides the editor with some very good clues about the relative weight of every concept in the category. This is even more feasible if the editor is presented with a ranking of candidates, ordering them by their estimated usefulness to describe the category. On the other hand, the process of abstracting a general concept from a list of weighted concepts is still difficult without previous knowledge about the collection, and some effort is required even from a trained editor. Here is a sample set of terms from our own research:

<p>Category: [1.1]</p> <p>Documents: n_3.txt n_8.txt n_14.txt n_6.txt n_12.txt n_13.txt n_19.txt n_7.txt n_9.txt n_2.txt n_10.txt n_11.txt</p> <p>Candidate Labels: guztiak 156.66631276494036 aukera 153.08784230524168 eta 121.82456157288982 list of candidates 68.9570404079377 public prosecutor 60.84444741876857 illegalization 44.61926144043027</p> <p>Sub-categories: {1.1.1} {1.1.2} {1.1.3}</p>

TABLE 2: Candidate labels for a cluster of documents relating to the illegalization of Aukera Guztiak, a political party related to a terrorist organization. Includes relative weights of each term to help editors

3.3 Fully automated category labelling

The automatic extraction of category labels has been attempted with different degrees of success by various researchers [11, 12, 13, 14, 15]. In essence, the system decides upon a single label (which might be a term or a phrase) that describes the category. Glover et al. [15] provide an interesting insight on category labelling. Terms are weighted by their relative importance to the topics of a category, but prior to that phase they are classified into three main types, parent nodes, child nodes and self nodes. This division is based upon the ratio of intra-cluster / collection frequency of every candidate label. This approach is highly interesting but suffers from the need of establishing a variety of thresholds for every collection. Results obtained by means of considering text surrounding hypertext anchors are quite good, but is only tractable in restricted environments or using huge resources.

Tonella et al. [14] create a ranking function for proposed terms or phrases. The ranking is mainly used for evaluation purposes, to assess the degree of accuracy of results. The system does not perform as well as Glover's, but evaluation methods are more sound. This is probably the most interesting issue, a systematic and founded system of evaluating results. This kind of evaluation was impossible in the first set of covered papers, simply because no ranking was made of the terms.

4. DESCRIPTION OF PROPOSED RESEARCH

User edited approaches usually present the user with a wide set of examples. This approach is analytic in the sense that it provides small and ideally related concepts from which the user abstracts a more general label. Using one of those concepts to represent the whole category might prove wrong, because labels should reflect the whole contents of the category.

The third option provides us with the idea of selecting only labels that show a synthetic view of the contents, and not only a small fraction of them. This is why we think that future research should be done following the last examples. Being able to identify the degree of generality and specificity of a candidate label with respect to the collection and a given category looks like a solid basis for detecting synthetic labels matching the category. In this sense, Glover's approach is quite promising, although we think that further research is to be done in some respects, including evaluation, effects of hierarchical relations (reported experiments dealt with a flat set of categories), and a comparison of results quality using different sources of data.

4.1 Sources of data

The main source of labels for category extraction is document text, although other sources can be found, such as labels from an already existing classification scheme, keywords assigned by authors, text surrounding anchors or metadata. Classification schemes are probably the worst choice, because they are rigid and don't fit well with every topic. We tried to use the UDC as a source for labels for a news collection, but it simply did not fit journalistic topics. The same would probably happen with very new topics, like the case of research literature. Other more flexible sources as thesauri are strongly biased towards analytic (specific) concepts, which we would probably want to avoid, as stated before. Nevertheless we still have some interesting sources to test as to prove their convenience for category labelling.

For what we know about Glover's experiments using some of the external data might be of great help. For scientific documents we can use both the abstract and keywords as special sources of data. A great deal of Web content has also some metadata describing the text of the documents. What all these sources have in common is that they "talk" even if they are a part of it. Differences observed in [12] and [15] describing documents tell us about the distinctive character of some of them. Sources describing the contents of a document do usually include synthetic concepts, more general approximations to the subject, although an important part of descriptions might arguably treat contents in a more analytic way. It would be highly interesting to prove the importance of the use of this variety of sources in the extraction of labels as opposed to the use of the full text of documents.

4.2 Weighting and selection of label candidates

Establishing ways of determining the relative importance of labels to describe the categories is obviously a key issue to be solved. We can use a wide array of weighting functions described in the references above, but no previous consistent work has been done (to the best of our knowledge) on this respect because no standard evaluation measures have been proposed or used over more than one or two different approaches. On the other hand, weight is not the only factor to be considered when selecting a label candidate.

As we have said before, candidate labels are to be extracted from any of the cited sources, and ranked according to their weight to the category, but even though we should be able to decide among the candidates taking into account some other aspects. The first of these is the exclusivity of labels regarding the analysed category. A label has to be rare enough in the collection to approximately describe only topics related to a particular category. At the same time a label has to be general enough to describe all the documents in a category at the same time. Finally, when clusters (categories) are hierarchically dependent, labels should reflect a visible hierarchical distinction between those categories. Mentioned fully automated labelling methods fail to show us this feature because flat sets of categories were used during the process. Preference should be given to one of the categories, or the label eliminated altogether.

It is often the case that no term or phrase describes accurately the contents of a category, or it is so low weighted that it does not even appear among the label candidates. This again points to the direction of using external sources of data. The problem is approached in [12] by means of using ontologies in conjunction with WordNet. When multiple candidate labels occur, they are mapped onto the ontology using WordNet. Using the structure of the ontology the relations among these candidates are found, and a more general concept is searched that represents them all. If the structure and dominion of the ontology suits the collection this approach is claimed to work very well. The problem is that we usually don't have an ontology for every collection, and in fact, creating an ontology "ex-profeso" for a given collection could prove far more complex and time consuming than manually creating both the classification scheme and the training examples for those categories.

These two problems are already present in the referenced literature, but are yet far from being solved. Glover's approach is again the most interesting way of dealing with exclusivity and generality problems, but is based upon thresholds, and thresholds are by no means easy to determine. It is also based on a special description of documents that relies on incoming links, and this is not always easy or even feasible. What we are trying to achieve is a method that is parameter free and relies only on document text, which is the only source of data being present in every case. The problem of lack of generality of terms present in the collection (or their low statistical weight) could be tackled using ontologies or any other source of concept representation that uses hierarchical relations. The problem is that these representations should also be created automatically. One possible way of dealing with the problem might come from the automatic thesaurus construction field.

5. RESEARCH METHODOLOGY AND PROPOSED EXPERIMENTS

As stated before, one of the major problems we see in cluster labelling research is the lack of standard evaluation measures. The most interesting works on this respect [11, 14] propose variations of precision and overlap measures based on the ranking of proposed labels and their match with manually assigned labels. The use of these measures over a wide range of different problems would provide us with fundamental information. Thus, this will be the first step to be taken.

We are currently working on providing a solid environment for experiments in this field. These experiments would include the comparison of label extraction procedures based on different sources for the same documents, different weighting functions and thresholds using only text derived directly from mentioned data sources and finally the use of automatically constructed thesauri for label subsumption.

Our first step has been the construction of a test collection. We used Cogprints, a repository of scientific papers on cognitive psychology and related areas. It includes some 2800 documents structured in a hierarchical way (of only 2 levels of depth, which is its major drawback) about several topics, including neural networks, biology, artificial intelligence, philosophy and linguistics. Almost every document has an abstract and keywords (or is excluded from the set), some metadata and the full text. This provides us with the possibility of exploring the effect of using different sources of data, and gives us a hierarchy and multiple topics to simulate a real situation. Every document is labelled according to one or more categories in the classification scheme, and groups of documents defined so manually created labels are always ready for comparison with the automatically extracted ones.

This collection can still be improved by providing synonyms or various correct answers to the problem, because direct match can be incredibly difficult to achieve if no external sources are present, although correction factors should be included in the evaluation measures to deal with this extra help.

6. DISCUSSION

It is still to determine the effect that NPL techniques could have over label creation, but would be really interesting to test them and find it out. Apart from natural language creation, other techniques as Named Entity Recognition or Part of Speech Tagging could be of high interest for our current research. POS tagging can lead us to have better probabilities of choosing the correct candidates, eliminating adjectives and verbs and selecting nouns and noun phrases as more proper candidates. NER could help us to devise auxiliary indexes, including onomastic, geographical and institutional indexes.

ACKNOWLEDGEMENTS

I would like to give special thanks to Vicente P. Guerrero-Bote from the UEX for his counsel and good ideas, especially in relation to evaluation issues.

REFERENCES

- [1] Chuang, Shui-Lung, and Lee-Feng Chien. "Automatic Query Taxonomy Generation for Information Retrieval Applications". In *OIR* 2003.
- [2] Chuang, Shui-Lung and Chien, Lee-Feng. "Taxonomy generation for text segments: A practical web-based approach". *ACM Trans. Inf. Syst.* 23(4): 363-396 (2005).
- [3] Müller, Adrian; Dörre, Jochen; Gerstl, Peter and Seiffert, Roland. "The TaxGen Framework: Automating the Generation of a Taxonomy for a Large Document Collection." In *HICSS*. 1999.
- [4] Suan, Melvyn. "Semi-Automatic Taxonomy for Efficient Information Searching." In: *Proceedings of the 2nd International Conference on Information Technology for Application (ICITA 2004)*. 2004, 130-136.
- [5] Hearst, Marti A., and Jan O. Pedersen. "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results." In: *Proceedings of the ACM SIGIR'96*. ACM Press, Zurich, 1996, 76-84.
- [6] Larocca Neto, J.; Santos, A.D.; Kaestner, C.A.A. and Freitas, A.A.. "Document Clustering and Text Summarization". In: *Proc. 4th International Conference Practical Applications of Knowledge Discovery and Data Mining (PADD-2000)*. London, 2000, The Practical Application Company. 41-55
- [7] Rauber, Andreas, and Dieter, Merkl. "Using Self-Organizing Maps to Organize Document Archives and to Characterize Subject Matter: How to Make a 342 Map Tell the News of the World." In *Database and Expert Systems Applications*. 1999, 302-311.

- [8] Rauber, Andreas. "LabelSOM: on the labeling of self-organizing maps". *Neural Networks*. Vol 5: 3527-3532 (1999).
- [9] Schwanke, R. W. and Platoff, M. A. 1989. Cross references are features. In: *Proceedings of the 2nd international Workshop on Software Configuration Management* (Princeton, New Jersey, United States, October 24 - 27, 1989). R. N. Taylor, Ed. ACM Press, New York, NY, 86-95.
- [10] Azcarraga, A.P. Ming-Huei Hsieh Pan, S.L. Setiono, R. "Extracting salient dimensions for automatic SOM labeling". In: *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*. Volume: 35, Issue: 4. 595- 600.
- [11] Hammouda, Khaled M.; Matute, Diego N. and Kamel, Mohamed S. "CorePhrase: Keyphrase Extraction for Document Clustering." *MLDM 2005*: 265-274
- [12] Halkidi, Maria; Nguyen, Benjamin; Varlamis, Iraklis; Vazirgiannis, Michalis: "THESUS: Organizing Web document collections based on link semantics." *VLDB J.* 12(4): 320-332 (2003).
- [13] Maqbool, O. and Babri, H.A. "Interpreting clustering results through cluster labeling." In: *Proc. of the IEEE Intl. Conference on Emerging Technologies*. 429-434. (September 2005).
- [14] Tonella, Paolo; Ricca, Filippo; Pianta, Emanuele and Girardi, Christian. "Using Keyword Extraction for Web Site Clustering", In: *Proc. of WSE 2003, 5th International Workshop on Web Site Evolution*, Amsterdam, The Netherlands, (September 2003).
- [15] Glover, Eric J.; Tsioutsoulouklis, Kostas; Lawrence, Steve; Pennock, David M. and Flake, Gary W. "Using Web structure for classifying and describing Web pages." In: *Proceedings of WWW-02, International Conference on the World Wide Web*. ACM Press, New York, US, Honolulu, US, 2002, 562–569.
- [16] Sebastiani, Fabrizio. "A Tutorial on Automated Text Categorisation" In: *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence*. (Analia Amandi, y Ricardo Zunino, eds.). Buenos Aires, AR, 1999, 7–35.
- [17] Gaussier, Éric; Goutte, Cyril; Popat, Kris; Chen, Francine. "A hierarchical model for clustering and categorising documents". In: *Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research* (Fabio Crestani, Mark Girolami, y Cornelis J. van Rijsbergen, eds.). Springer Verlag, Heidelberg, DE, Glasgow, UK, 2002, 229–247. Published in the "Lecture Notes in Computer Science" series, number 2291.
- [18] Steinbach, Michael; Karpys, George; Kumar, Vipin. "A comparison of document clustering techniques". In: *KDD Workshop on Text Mining*, 2000.
- [19] van Rijsbergen, Keith. *Information Retrieval*. Rutterworths, London, 1979.
- [20] Manning, Christopher and Schüze, Hinrich. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, US, 1999.