A proposal for building domain topic taxonomies

Maria Fernanda Moura[1,2], Ricardo Marcondes Marcacini[1], Bruno Magalhães Nogueira[1], Merley da Silva Conrado[1] and Solange Oliveira Rezende[2]

[1] CNPTIA/Embrapa http://www.cnptia.embrapa.br - fernanda@cnptia.embrapa.br
[2] ICMC/USP http://www.icmc.usp.br - marcacini@grad.icmc.usp.br, {brunomn,merleyc,solange}@icmc.usp.br

In this work a methodology to aid the process of organizing text collections is proposed, aiming to reflect exactly the existent and recoverable publications in a specific domain [3]. Although the methodology is completely automatized, it allows the domain specialist to intervene in some of its steps; in this way, some of the comprehensibility aspects can be satisfied when necessary. The methodological basis is a text mining process, which uses statistical machine learning models, specially exploratory and multivariate data analysis. The goals of the process are to identify the domain vocabulary presented in the text collection, grouping the text collection in hierarchical clusters and automatically generating a topic taxonomy. The obtained topic taxonomy can be edited by the final users oriented by statistical measures. Additionally, all the computational support is being developed in a way this permits the incorporation of new methods, when necessary or as an alternative to the methods already incorporated. The methodology process can automatically decide which methods to use in some steps or can guide the users in their decisions. The process steps are illustrated in Fig. 1.

In the problem identification step, the scope and goals of the methodology ap-

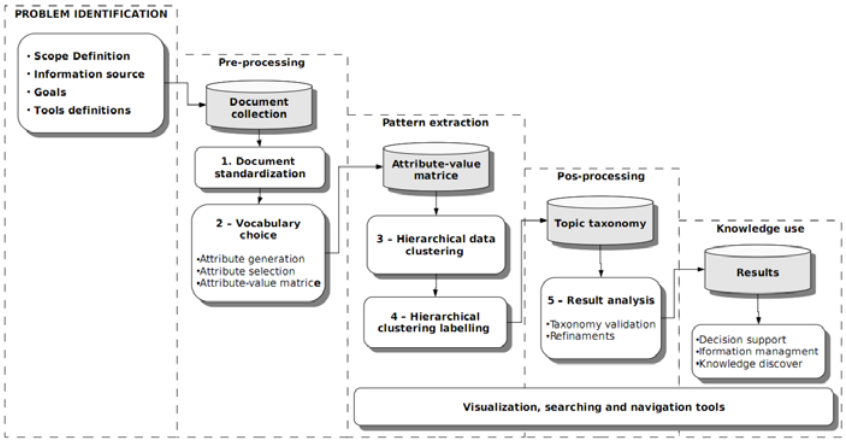

Fig. 1. Topic taxonomy methodology process

plication are defined. In the preprocessing, the documents are standardized, the

interesting attributes are found and the attribute space dimension is reduced. This is a complex step and it is the keyprocess, because the quality of the all obtained results is going to depend on it. At the end of this step an attribute-value matrice can be constructed. In the pattern extraction step, a hierarchical cluster algorithm is used and the resulting groups are automatically labeled [4]. A label is a term or a set of terms that can discriminate the cluster. This cluster labelling process generates a taxonomy in which there are not term repetitions along the labels in the same hierarchical branch; because a specific node has its specific label and it can have children that inherit its label. The labelling method does an objective analysis of all the terms selecting those that are discriminative for a cluster and eliminating those that do not contribute to discriminate any cluster. An objective and a subjective evaluation of the generated taxonomy are done in the pos-processing step. If the generated taxonomy is not satisfactory enough, the user can try to identify some criteria to return to other parts of the process. On the other hand, the taxonomy can be refined. The domain specialists can edit and evaluate their modifications over the taxonomy aided by statistical measures. Additionally, they can create new terms based on the taxonomy terms. Finally, the taxonomy can be used to facilitate the process of organizing, understanding and retrieving information. Moreover, the obtained taxonomy can be used as decision support data, because the topics represent different tendencies. The proposed methodology has been used in production experiments with specific goals, as the case studies of scientifc production tendencies in beef and dary cattle along with Embrapa Pecuária Sudeste [2] and a digital library organization [1]. For the research, artifical intelligence text collections has been used for prospection and validation of new methods due to the proximity of domain specialists in this area with the project [3].

The authors would like to thank CAPES, CNPq and IFM for the technical and finance support.

References

1. R. M. Marcacini, M. F. Moura, and S. O. Rezende. Biblioteca digital do ifm: uma aplicação para a organização da informação através de agrupamentos hierárquicos. In *XIII Brazilian Symposium on Multimedia and the Web - WDL 2007 - III Workshop on Digital Libraries*, volume CD-ROM of *WDL*, pages 1–16, 2007.
2. C. A. Mazzari. Gestão de pessoas e identificação de competências estratégicas em unidades descentralizadas da embrapa - o caso embrapa pecuária sudeste, 2007. Research and Development Project, Status: active, Research center: Embrapa Pecuária Sudeste (http://www.cppse.embrapa.br).
3. M. F. Moura, R. M. Marcacini, B. M. Nogueira, M. S. Conrado, and S. O. Rezende. Uma abordagem completa para a construção de taxonomias de tópicos em um domínio. Technical Report 329, ICMC/USP, 2008.
4. M. F. Moura and S. O. Rezende. Choosing a hierarchical cluster labelling method for a specific domain document collection. In *New Trends in Artificial Intelligence.*, chapter 11, pages 812–823. EPIA- Encontro Portugues de Inteligência Artificial, Guimarães, Portugal, 1 edition, 2007.

[3] Text collections are available at http:www.labic.projects.