TaXEm: a tool for aiding the evaluation of domain topic

Merley S. Conrado[1] and Maria Fernanda Moura[2] and Solange O. Rezende[1]

[1] Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)

[2] Embrapa Informática Agropecuária

{merleyc,solange}@icmc.usp.br, fernanda@cnptia.embrapa.br

The notorious advances in textual information storage need fast and efficient tools to organize, retrieve and browse this information and tools for knowledge extraction. A very interesting way to organize specific domain information is the topic taxonomy building. Moreover a great challenge in this research area is the result evaluation and validation. This evaluation can be carried out through objective measures or through a subjective analysis, which is based on the domain specialist judgment. The measures CQM and SQM [1] are used to evaluate a generated taxonomy against a reference taxonomy. A reference taxonomy is constructed by human and consolidated through its community use along the years. The CQM is used to evaluate the generated taxonomy relatively to the selected descriptors for each taxonomy node; on the other hand, the SQM is used to evaluate the taxonomy structure. As these objective measures do not encompass the specialist knowledge, the specialist evaluation is very important. However the human evaluation is expensive, because this task involves readiness, time and dedication from the specialists. In this way, the TaXEm tool claims to reduce the subjective evaluation costs. The TaXEm (Taxonomia em XML da Embrapa) tool offers subsidies for carrying out a taxonomy (semi)automatic evaluation, which allows the user to implement some automatic evaluation before going on a subjective evaluation. Initially the TaXEm development was based on the Embrapa´s Information Agency[3] considering each product agency [2] as a reference taxonomy; although it can be configured to work with any other reference topic taxonomy. In this work, a topic taxonomy is a hierarchical organization of a text collection, in which each node has a descriptor set of terms and each document can have its own set of descriptors.

The TaXEm implements an enrichment process from the reference taxonomy and its related set of documents. This enrichment process has its essential character in the descriptor term set expansion, taking the original set of terms and finding equivalent terms, synonyms or semantically related terms in a domain thesaurus. The descriptor term set can be keywords, subject, category, title or any other metadatum which was manually attributed to each reference taxonomy node or to each related document. The original descriptor terms with the expanded terms form the expanded vocabulary. For example, the Embrapa´s Information Agency belongs to the agribusiness domain, consequen-

[3] Embrapa Information Agency - http://www.agencia.cnptia.embrapa.br/

tly a agriculture thesaurus was used to expand the vocabulary - THESAGRO[4]. For example, in this thesaurus, the term abacaxizeiro (pineaple tropical plant) has an equivalent relation to abacaxi (pineapple fruit) and the term enologia (enology) has a semantically relation to vinho (wine). If a medical topic taxonomy were constructed, for example, using some topic in the PubMed collection (http://www.ncbi.nlm.nih.gov/sites/entrez/), the TaXEm could use the MeSH thesaurus (http://www.ncbi.nlm.nih.gov/mesh). In this second example, a term as adipose tissue neoplasm is a semantic generalization for angiolipoma, angiomyolipoma, lipoma, liposarcoma and myelolipoma.

In this way, the TaXEm allows the comparison of an automatic generated topic taxonomy against an enriched reference taxonomy without information loss.

As the tool generates expanded vocabularies, the verifications among the selected descriptor terms can be done at the document set level or the taxonomy hierarchy level. This verification consists in the calculation of the proportion of the selected terms that belongs to the set of the enriched terms in each document or taxonomy level. It has to be noted that the enrichment process increases the verification space which can increase the proportion of corrected matches. Moreover, it is possible to verify if the descriptor terms are able to discriminate the domain topics and subtopics through the verification of their position in each hierarchy node or the position of the nodes into the hierarchy structure. Additionally, the domain documents were organized in textual bases and their vocabulary has been enriched, so that a variety of classification tasks could be performed over them. To sum up, the expansion of the vocabulary has contributed with an objective evaluation of the generated taxonomies and has enabled a tool to transform this kind of evaluation in an automatic or semi-automatic process.

The authors thank the support from CNPq and EMBRAPA to this work.

Referências

1. V. Kashyap, C. Ramakrishnan, C. Thomas, and A. Sheth. Taxaminer: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, 1(2):240–266, 2005.
2. M. I. F. Souza, A. D. d. Santos, M. F. Moura, and M. d. D. R. Alvez. Agência de informação Embrapa: uma aplicação para a organização da informação e gestão do conhecimento. In *XX Simpósio Brasileiro de Engenharia de Software (SBES) & XXI Simpósio Brasileiro de Bancos de Dados (SBBD) - II Workshop de Bibliotecas Digitais (WDL)*, pages 51–56, 2006.

[4] http://www.agricultura.gov.br/portal/page?_pageid=33,959135&_dad=portal&_schema=PORTAL