FACILITANDO A AVALIAÇÃO DE TAXONOMIAS DE TÓPICOS AUTOMATICAMENTE GERADAS NO DOMÍNIO DO AGRONEGÓCIO

MERLEY DA SILVA CONRADO[1]
MARIA FERNANDA MOURA[2]
SOLANGE OLIVEIRA REZENDE[3]

RESUMO: O custo de avaliação de taxonomias automaticamente geradas costuma ser alto se realizado exclusivamente por especialistas do domínio. Neste trabalho propõe-se uma forma de validação objetiva para o vocabulário automaticamente obtido, com o qual uma taxonomia automaticamente gerada é representada. Para isso, o vocabulário gerado é comparado ao de uma taxonomia pré-existente, já validada, bem aceita e que use um thesaurus ou dicionário específico do domínio. Como taxonomias validadas foram utilizadas algumas árvores da Agência de Informação Embrapa e como vocabulário padrão o contido no Thesagro.

PALAVRAS-CHAVE: taxonomias de tópicos, vocabulário controlado, extração de termos

MAKING THE EVALUATION OF AUTOMATIC GENERATED TOPIC TAXONOMIES EASIER IN THE AGRIBUSINESS DOMAIN

ABSTRACT: The subjective evaluation of automatic generated topic taxonomies use to be expensive when is done exclusively by domain specialists. In this paper, we propose an objective way to evaluate the automatically generated vocabulary that represents an automatically generated taxonomy. To reach this goal, the generated vocabulary is compared to the one used in a gold taxonomy, that means a validated pre-existent taxonomy which had used terms from a thesaurus in its definitions. As gold taxonomies, some knowledge trees from EMBRAPA's Information Agency were used, as well as the vocabulary presented in the Thesagro (the Brazilian thesaurus for agriculture).

KEY-WORDS: topic taxonomies, controled vocabulary, term extraction.

1. INTRODUÇÃO

Estudos indicam que, aproximadamente 80% de toda a informação das corporações no mundo é representada por documentos textuais[4], dado que essa é a forma mais natural de armazenar informações. Esse elevado volume de dados e informações faz com que seja necessário sua organização de forma mais rápida e eficiente; o que seria um trabalho praticamente impossível de se fazer manualmente. Tal organização traz, em geral, vantagem competitiva para as empresas, pois mapeia o conhecimento explícito, possibilitando a recuperação de informação e a descoberta de novos conhecimentos para suporte à tomada de decisões.

Frente a esse cenário, a gestão do conhecimento, no que tange à organização do conhecimento explícito e à elaboração de mapas de conhecimento em um domínio, pode ser apoiada pela mineração de textos com a construção de taxonomias de tópicos de forma semi-automática ou automática (Moura et al., 2008). Um dos grandes problemas nesse processo, além da própria construção dos tópicos, é a validação dos resultados obtidos. Uma boa solução é comparar de

[1]Mestranda, USP – São Carlos, E-mail: merleyc@icmc.usp.br
[2]Pesquisadora e doutoranda, Embrapa Informática Agropecuária e USP – São Carlos, E-mail: fernanda@cnptia.empbra.br
[3] Docente, Departamento de Ciências de Computação, ICMC-USP – São Carlos, E-mail: solange@icmc.usp.br
[4] Delphi Group - http://www.delphigroup.com/

forma objetiva uma taxonomia consolidada com a gerada de forma automática, com o auxílio de especialistas da área de conhecimento mapeada. Porém, é uma solução cara, dado que envolve tempo e dedicação desses especialistas.
Neste sentido, este trabalho visa diminuir o custo da avaliação subjetiva, economizando-lhe algumas etapas. A etapa economizada, no escopo deste trabalho diz respeito exclusivamente à validação do vocabulário escolhido para representar as taxonomias automaticamente geradas.
Foram escolhidas, para caso de uso, algumas árvores publicadas pela Agência de Informação EMBRAPA como taxonomias *gold*, isto é, taxonomias padrão para comparação. E, no escopo deste trabalho, é apresentada a preparação de dados das Agências de Informação Embrapa para o padrão necessário à comparação com o vocabulário automaticamente obtido e respectivas taxonomias de tópicos automaticamente geradas.

2. A Taxonomia de Tópicos da Agência de Informação EMBRAPA

Uma das formas de divulgação de informação na EMBRAPA é a Agência de Informação, que "é um sistema Web que possibilita a organização, o tratamento, o armazenamento, a divulgação e o acesso à informação tecnológica e ao conhecimento gerados pela Embrapa e outras instituições de pesquisa"[5] (Souza et al, 2006). Os dados e informações contidos na Agência são organizados em hierarquias, denominadas árvores do conhecimento. Nos primeiros níveis das hierarquias estão os conhecimentos mais genéricos e, nos níveis mais profundos, estão os conhecimentos específicos (Evangelista et al., 2003). Cada nó das hierarquias corresponde a um tema (tópico) descrito por um texto, que resulta da compilação do conhecimento produzido por pesquisadores, técnicos extensionistas e agricultores; e, ainda faz referências a outras obras que complementam essa informação. Tais hierarquias são visualizadas por meio de árvores hiperbólicas ou navegação por hipertexto das páginas Web. Adicionalmente, são fornecidas ferramentas de busca alimentadas por palavras-chave.
Utilizar árvores da Agência de Informação Embrapa como taxonomias *gold* é interessante, pois a informação é pública, consolidada, validada e completa. Mesmo assim, algumas adaptações foram necessárias, devido às características das taxonomias envolvidas no processo, pois a Agência é composta por árvores manual e subjetivamente construídas; e tais árvores estão sendo comparadas a taxonomias automaticamente geradas.
A primeira característica é que para cada nó, em uma taxonomia de tópicos automaticamente obtida, o tópico possui um ou mais termos como palavras-chave a serem validadas, e cada um desses termos possuem similares, que poderiam ter sido também encontrados. Na árvore de conhecimento da Agência, cada nó é representado apenas pelos termos mais específicos, que foram subjetivamente selecionados a partir de um thesaurus. Por exemplo, o termo *Abacaxi*, neste domínio, poderia também ser representado pelo termo *Ananas Comosus*, assim, o nó deveria conter os termos *Abacaxi* e *Ananas Comosus*. Dessa forma para comparar a taxonomia automaticamente gerada com uma *gold*, sem perda de informação, foi necessário expandir os termos em cada nó das árvores da Agência, buscando termos sinônimos e/ou relacionados no thesaurus utilizado, no caso, o Thesagro[6] (Thesaurus Nacional Agrícola).
A segunda característica é que, taxonomias automaticamente obtidas agrupam documentos em tópicos, e, a seguir procuram palavras-chave para descrever os tópicos nesses documentos. Na Agência, em cada nó da hierarquia há uma página Web contendo uma síntese assunto do nó em questão e, possivelmente mais documentos relacionados ao nó são referenciados, bem como metadados desses documentos. Novamente, uma adaptação faz-se necessária, a fim de abstrair esses textos como documentos agrupados sob o nó.

---

[5] Agência de Informação Embrapa - http://www.agencia.cnptia.embrapa.br/
6 Thesagro - http://www.agricultura.gov.br/portal/page?_pageid=33,959135&_dad=portal&_schema=PORTAL

3. Adaptação da Agência a uma Taxonomia *Gold*

O processo de obtenção automática de taxonomias de tópicos necessita da extração de termos capazes de representar com qualidade cada nó. Além disso, esses termos podem e devem ser utilizados como palavras-chave dos documentos, podendo alimentar expressões de busca na coleção de textos ou em coleções similares. Deve-se ressaltar que, não necessariamente, esses termos representam adequadamente o domínio de conhecimento, pois são obtidos a partir de uma coleção restrita de documentos. E, ainda, espera-se que o volume de termos extraídos automaticamente da coleção seja maior que um conjunto de termos selecionados por humanos em um thesaurus. Assim, para avaliar a qualidade dos termos extraídos das mesmas coleções de documentos que geraram algumas árvores da Agência, implementaram-se os processos de adaptação de alguma árvore da Agência a uma taxonomia *gold*. Para efetuar as adaptações foi desenvolvida a ferramenta TaXEm (Taxonomia em XML da Embrapa), que utiliza como base de vocabulário controlado o contido no Thesagro. Deve-se ressaltar que, apenas Agências de produto foram utilizadas, a fim de garantir que o thesaurus utilizado fosse o Thesagro; dado que essas árvores são baseadas em cadeia produtiva de produtos agrícolas.

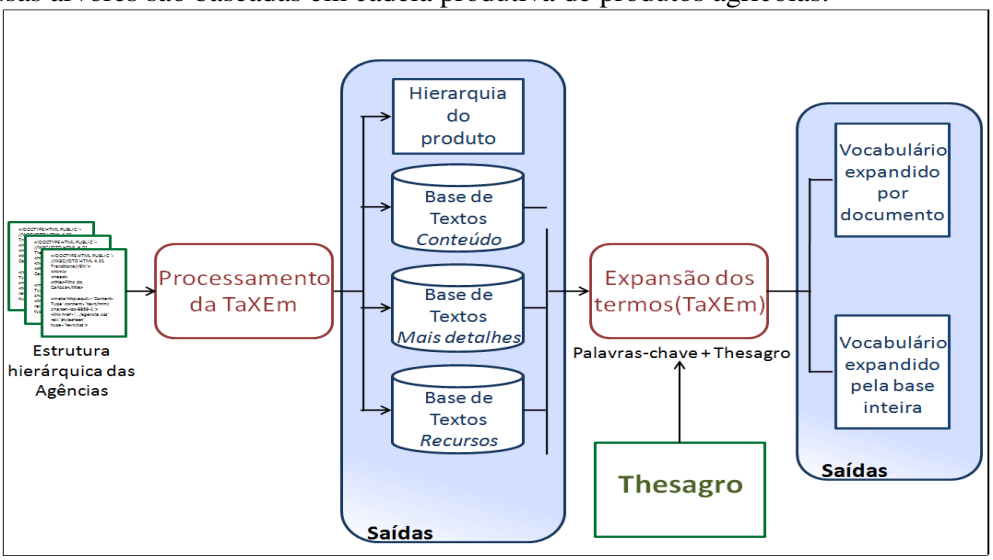
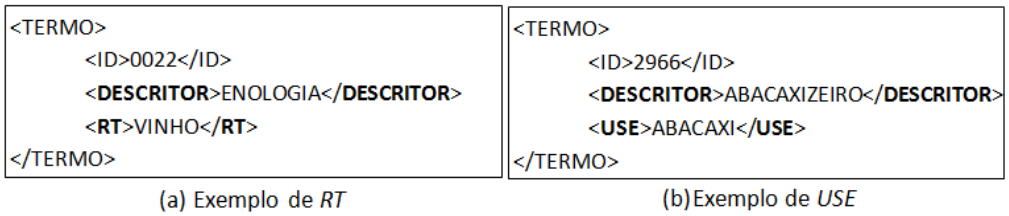


Figura 1: Processo para a preparação do conteúdo das Agências

Conforme mostrado na Figura 1, o processo de adaptação proposto neste trabalho consiste em: a partir dos arquivos que armazenam a estrutura hierárquica das Agências, reestruturá-la de modo que seja possível facilmente vincular os documentos relacionados a cada nó; e que, cada nó seja rotulado com os termos que o representam, viabilizando uma busca mais ampla.
Assim, para cada entrada, a hierarquia dos nós para o produto é preparada em formato *.xml*, e em cada nó, os documentos relativos ao mesmo são referenciados no *xml*. Estes documentos são gerados em formato de texto plano, separadamente da hierarquia, em três diferentes categorias: *conteúdo*, *mais detalhes* e *recursos*. Os documentos pertencentes à categoria *conteúdo* descrevem o tópico, nos moldes da Agência; os documentos da categoria *mais detalhes* são os metadados sobre os recursos; e os pertencentes a *recursos* são os próprios recursos disponíveis do nó, isto é, as informações complementares citadas no texto que descreve o nó.
A seguir, o processo permite expandir os termos originais dessa hierarquia, ou seja, com a utilização do Thesagro aumentar as possibilidades de cada palavra-chave acrescentando-lhes sinônimos ou termos relacionados. Utilizando a notação do Thesagro, na expansão de cada termo *T* (descritor), consideraram-se os termos relacionados (*RT*) ao termo buscado *T*; sendo

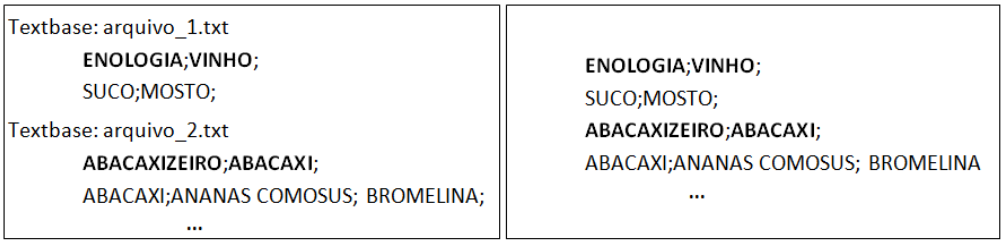
que, os *RT* possuem significados que se relacionam semanticamente com o termo *T*, mas sem nenhuma ligação hierárquica entre si. Também são considerados os termos *USE,* que são classificados, neste trabalho de expansão, como tendo uma relação de equivalência ao termo *T*. Na Figura 2 são mostrados exemplos de (a) *RT* e (b) *USE.*



(a) Exemplo de *RT* (b) Exemplo de *USE*

Figura 2: Exemplos de ***RT*** e ***USE***

O vocabulário expandido, obtido com a utilização do Thesagro, é apresentado de duas formas: (a) separado por documento e (b) referente à base inteira, conforme ilustrado na Figura 3.



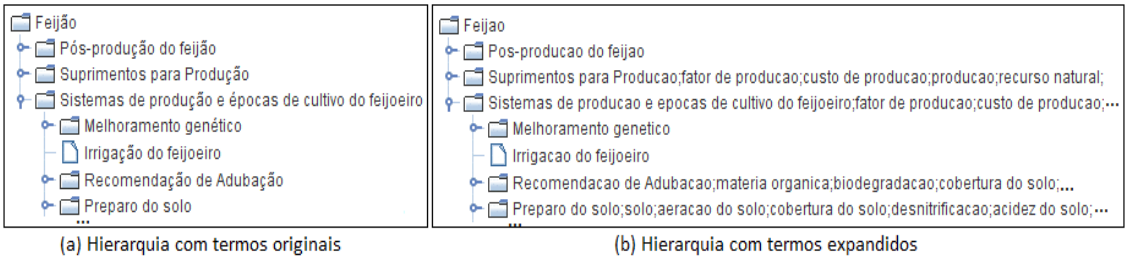
(a) Exemplo por documento (b) Exemplo pela base inteira

Figura 3: Vocabulário expandido

4. AVALIAÇÃO DOS TERMOS EXTRAÍDOS DOS DOCUMENTOS

A primeira forma de avaliação é a verificação de acertos entre os termos extraídos e os termos originais, o que permite conferir se os termos extraídos contêm termos importantes e representativos desse domínio. Esta verificação é feita a partir do vocabulário expandido obtido da TaXEm contra o automaticamente obtido. Com a expansão, a forma de descrever os termos é ampliada, aumentando, portanto, o espaço de verificação de acertos. Utilizando os resultados produzidos pela TaXEm, esta verificação pode ser feita em cada documento ou para a base como um todo, dependendo do objetivo pré-estabelecido.

A segunda forma de avaliação considera como base a taxonomia *gold*, com o objetivo de verificar se os termos extraídos representam os tópicos e sub-tópicos do domínio. Dessa forma as comparações podem ser feitas verificando os termos de cada nó em cada hierarquia ou verificando a posição dos nós das hierarquias. Isso é possível devido ao padrão da hierarquia preparada na TaXEm, o que permite sua fácil visualização, utilizando, por exemplo, a ferramenta TaxTools (Marcacini e Rezende, 2008), como ilustrada na Figura 4 que usa como exemplo o produto *Feijão,* conforme publicado na Agência Embrapa. Nesta figura é mostrada (a) a hierarquia da árvore do *Feijão* reestruturada pela TaXEm contendo somente os termos originais e (b) a mesma hierarquia contendo os termos expandidos com o vocabulário.



(a) Hierarquia com termos originais (b) Hierarquia com termos expandidos

Figura 4: Exemplo da visualização da hierarquia

6. CONSIDERAÇÕES FINAIS

A busca por uma forma de avaliar os termos automaticamente extraídos de uma coleção de textos em um domínio tem como objetivo economizar o trabalho do especialista de domínio, ao avaliar uma taxonomia de tópicos automaticamente gerada. Assim, neste trabalho, procurou-se pela adaptação de taxonomias existentes a uma forma padrão para comparação automática. Desse modo, acredita-se que a expansão do vocabulário e as formas de validação aqui propostas contribuem para a tarefa de validação de duas formas: obtém-se um método objetivo, que pode dar suporte às ferramentas automáticas, independentemente da análise subjetiva; e economiza etapas em uma análise subjetiva, deixando para o especialista do domínio as tarefas que, de fato, demandam seu conhecimento.

Como trabalho futuro, estão em implementação, sob uma única ferramenta, algoritmos de comparação automática entre os vocabulários obtidos e a *gold*, e algoritmos de comparação entre as taxonomias obtidas e a *gold*. Os dois processos de comparação utilizam as expansões realizadas pela TaxEM, dado que a validação visual dos resultados permitiu completar as definições e seus algoritmos.

6. REFERÊNCIAS

SOUZA, M. I. F.; SANTOS, A. D. dos; MOURA, M. F.; ALVES, M. das D. R. Agência de informação Embrapa: uma aplicação para a organização da informação e gestão do conhecimento. In: Simpósio Brasileiro de Engenharia de Software, 20.; Simpósio Brasileiro de Bancos de Dados, 21.; **Workshop de Bibliotecas Digitais**, 2., 2006, Florianópolis. Anais... Florianópolis: SBC, 2006. p. 51-56.

EVANGELISTA, S. R. M.; SOUZA, K. X. S.; SOUZA, M. I. F.; CRUZ, S. A. B.; LEITE, M. A. A.; SANTOS, A. D.; MOURA, M. F. Gerenciador de conteúdos da Agência Embrapa de Informação, Curitiba, C. P. U. C. **International Symposium on Knowledge Management (ISKM)**, v. 1, p. 1-12, 2003.

MARCACINI, R. M.; REZENDE, S. O. Técnicas de Visualização de Informação para Análise de Taxonomias de Tópicos Hierarquicamente Relacionados. In: **II Assembléia Geral do IFM**, Campinas, SP. Instituto Fábrica do Milênio, São Carlos, SP, v. 1, p. 1-248, 2008.

MOURA, M. F.; MARCACINI, R. M.; NOGUEIRA, B. M.; CONRADO, M. S.; REZENDE, S. O. A proposal for building domain topic taxonomies. In: **I Workshop on Web and Text Intelligence (WTI) - XIX Simpósio Brasileiro de Inteligência Artificial (SBIA)**, Salvador, BA. Proceedings of I Workshop on Web and Text Intelligence, São Carlos, SP, ICMC/USP, v. 1, p. 83-84, 2008.