CORRETOR ORTOGRÁFICO AUTOMÁTICO DE TERMOS THESAGRO PARA A AGÊNCIA DE INFORMAÇÃO EMBRAPA

Ribeiro, Rafael S.[1] (IC); Oliveira, Leandro H. M.[2] (CO)
rafael_pucc@hotmail.com
[1] *Departamento de Engenharia de Computação, Pontifícia Universidade Católica;*
[2] *Embrapa Informática Agropecuária.*

A Agência de Informação Embrapa[1] 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. Essas informações estão organizadas em uma estrutura em forma de árvore hiperbólica, denominada árvore do conhecimento, na qual o conhecimento é organizado de forma hierárquica. O controle e o gerenciamento dos conteúdos das Árvores de Conhecimento são realizados pelo sistema gestor de conteúdo da agência, um ambiente *web* para gestão da informação tecnológica que reúne um conjunto de ferramentas para o tratamento e a recuperação da informação. No conjunto de informações contidas nas árvores de conhecimento estão os recursos eletrônicos (arquivos de áudio, vídeo, imagens, textos e documentos), que são associados a conteúdos específicos das árvores para fornecer informações complementares. Os metadados dos recursos eletrônicos são descritos de acordo com o padrão *Dublin Core*. Especificamente, em relação ao assunto, adota-se o **Thesagro**, um *thesaurus* de termos agrícolas da Biblioteca Nacional de Agricultura sendo o mesmo eficaz para efetivação do controle terminológico, uma vez que disponibiliza um vocabulário controlado, apresentando estruturas de relação de equivalência, hierarquia e associação entre os termos contemplados. Para o fomento destes termos, a Embrapa Informática Agropecuária recebe eletronicamente um arquivo em formato *XML* que correspona uma parte do **Thesagro**. Tal arquivo possui termos com grafia incorreta (acentos e cedilha) que devem ser corrigidos. Este artigo apresenta um corretor ortográfico desenvolvido para corrigir automaticamente estes termos por meio do software *Aspell*[2], um corretor sintático de uso livre que apresentou os melhores resultados após a comparação entre vários corretores disponíveis gratuitamente na internet. O primeiro passo do corretor implementado é retirar todos os termos do arquivo enviado, mantendo as *TAGs*do *XML*intactas. Em seguida, usando o corretor *Aspell*, cada termo selecionado é verificado quanto a sua corretude ortográfica. Neste passo podem ocorrer duas possibilidades: 1) o termo é julgado como correto e não necessita de correções ou 2) o termo é julgado como incorreto e o sistema sugere uma correção. Após a correção, os termos são novamente inseridos no arquivo original no mesmo local onde estavam anteriormente. A quarta fase consiste na validação da correção automática. Para isso, os termos corrigidos na segunda etapa são comparados com os termos originais, gerando dois arquivos como resultados: um contendo a lista de termos corrigidos e outro contendo uma lista de termos não encontrados ou corrigidos por engano pelo corretor *Aspell*(como os termos originais possuem apenas erros de acentuação, caso alguma letra tenha sido alterada o termo é considerado como errado), que deverá ser verificada manualmente. Os resultados mostram uma considerável redução do trabalho humano de correções do termos, já que apenas uma pequena parte do arquivo original (cerca de 10%) deverá ser verificado.

Embrapa Informática Agropecuária

[1]http://www.agencia.cnptia.embrapa.br
[2]http://aspell.net/