Extração automática de candidatos a termos para criação de um mapa conceitual do domínio de Intensificação Agropecuária.

Felipe Heidi Shiratori, estudante de Engenharia de Computação, estagiário do Laboratório de Organização e Tratamento da Informação Eletrônica –
Leandro Henrique de Mendonça Oliveira, supervisor

V Mostra de Trabalhos de Estagiários e Bolsistas
Campinas, SP – 26 a 30 de outubro de 2009

Introdução

A mineração de textos em conjunto com a análise terminológica tem-se mostrado extremamente úteis para a organização e estruturação de conhecimento. Neste contexto, a evidência, o tratamento e a extração do conhecimento específico fornecida pelas técnicas de mineração textual e extração automática de termos servem de base para criação de mapas conceituais e ontologias de um determinado domínio do conhecimento.

Objetivos

O objetivo deste trabalho é extrair automaticamente os candidatos a termos de um *córpus* do domínio Intensificação Agropecuária, para subsidiar e facilitar a criação de um mapa conceitual deste domínio.

Materiais e Métodos

A compilação e limpeza dos textos que compõe o *córpus* textual foi a primeira tarefa a ser realizada, sendo formado por 398 textos num total de 2,5 milhões de palavras. A partir do *córpus*, os programas do Pacote NSP (*Ngram Statistics Package*) foram utilizados para realizar a extração automática dos candidatos a termos formados por listas de unigramas, bigramas e trigramas. Tais listas foram repassadas para diferentes especialistas do domínio para subsidiar a criação e estruturação do mapa conceitual almejado. O esquema deste processo pode ser visto na Figura 1.

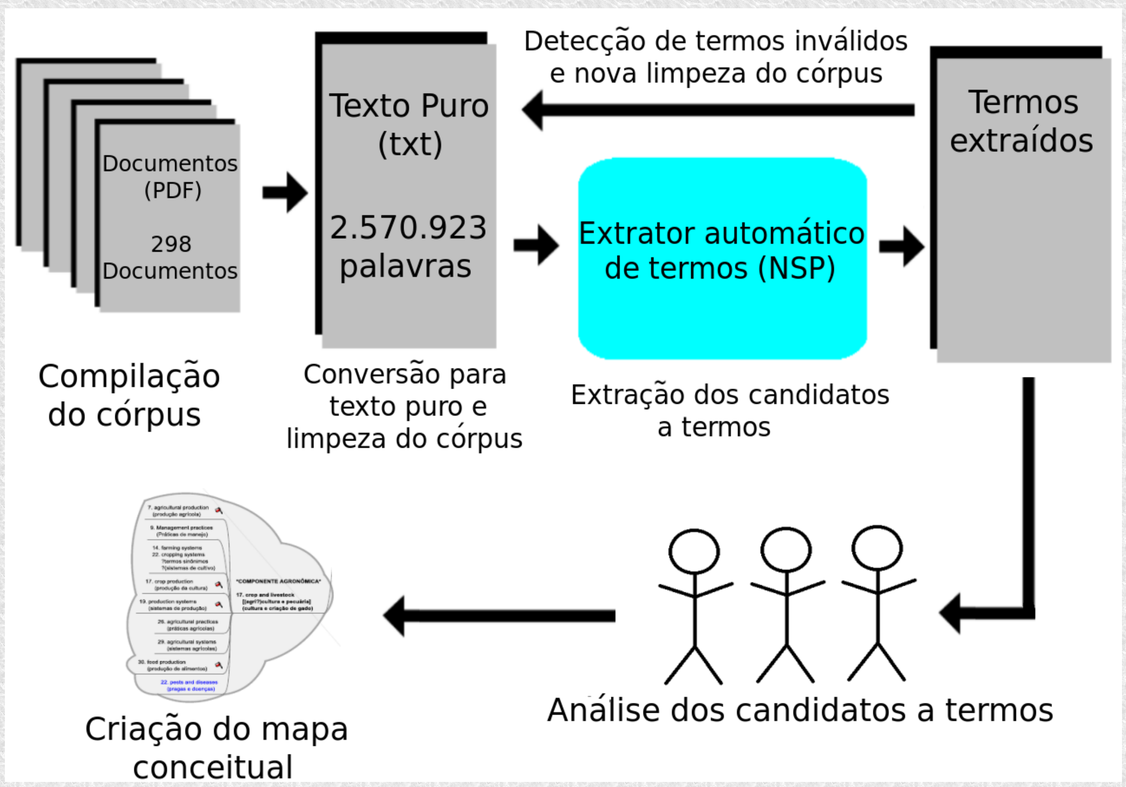

Figura 1. Diagrama das etapas da extração automática de termos

Como seguimento deste trabalho, para facilitar a tarefa de extração automática de termos, uma interface *web* das ferramentas disponíveis no Pacote NSP está sendo desenvolvida, utilizando um *Framework* AJAX em conjunto com a linguagem PHP.

Resultados e Discussões

Além do *córpus* textual, foram extraídos no total 6630 unigramas, 1687 bigramas e 2781 trigramas. Uma versão preliminar do mapa conceitual pode ser visto na Figura 2.

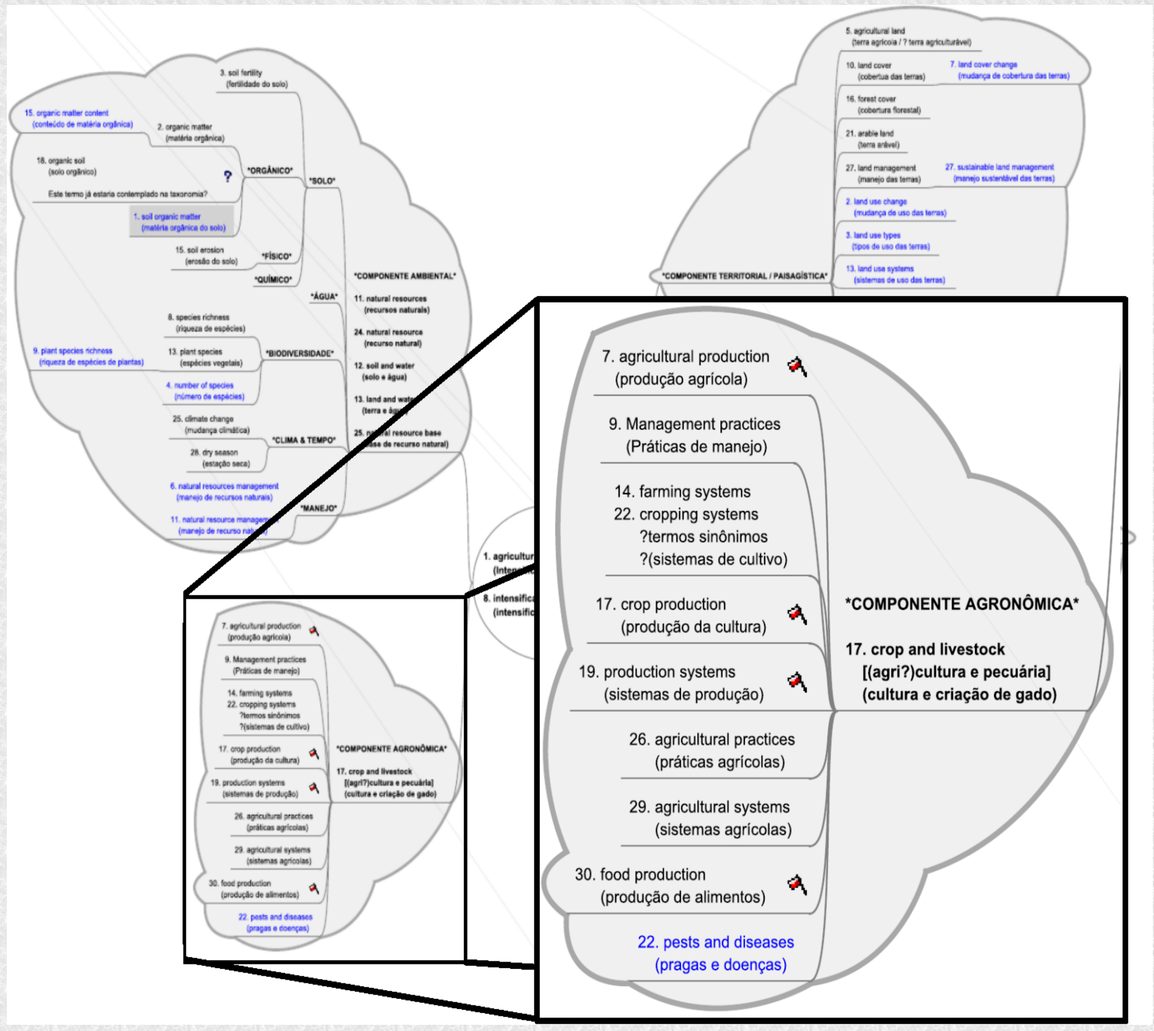

Figura 2. Mapa conceitual preliminar

Conclusões

Os resultados iniciais mostram que a extração automática de candidatos a termos foi extremamente útil, poupando o trabalho custoso de identificação manual dos termos, e servindo como subsídio fundamental para a criação do mapa conceitual.

Referências Bibliográficas

[1] *Ngram Statistics Package* http://www.d.umn.edu/~tpederse/nsp.html (acessado em Setembro/2009)

Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento