

DETECÇÃO AUTOMÁTICA DE SENTENÇAS DE ARTIGOS TÉCNICOS PARA EXTENSÃO DE ONTOLOGIA NÚCLEO

¹ **MISOKAMI, P. E. O.** (paulo_misokami@hotmail.com); ² **BATISTA JUNIOR, J.** (joinvile@ufgd.edu.br);

¹ Aluno do curso de Sistemas de Informação - UFGD; ² Professor da área de Computação da FACET/UFGD;

Artigos técnicos no formato PDF são abundantes na internet. A extração de informações de artigos técnicos é de grande valia para a extensão de ontologias que possam ser utilizadas para caracterização da evolução tecnológica em uma determinada área do conhecimento. Ontologias representam o conhecimento em um dado domínio a partir de relações entre conceitos, extraídas de textos no formato TXT. A extração automática de relações, a partir de textos de linguagem natural, é uma atividade essencial para a construção e extensão de ontologias. Os artigos utilizam duas colunas, de maneira geral, e a conversão para o formato TXT mistura nomes de seções, partes de sentenças distintas, notas de rodapé, etc. Os detectores de sentenças existentes não suportam a separação das sentenças das seções. Para o processamento de vários artigos técnicos é desejável um processo completamente automatizado, o que aumenta significativamente a complexidade da separação das informações. A metodologia utilizada se baseia na concepção do processo e da sua prototipação: (a) para situações distintas de representação de autoria; (b) separação das colunas do artigo e geração de uma coluna única; (c) separação de títulos de seções, sentenças das seções, figuras, notas de rodapé e referências; e (d) principais representação de referências. Os processos de separação de autoria, referências, e de títulos e sentenças de seções, notas de rodapé e textos descritivos de figuras foram identificados e protipados. Como resultado da prototipagem, da solução proposta neste trabalho, foi realizada a extração automática das várias partes componentes de um artigo técnico de prova: (a) nomes, e-mails e origem dos cabeçalhos de autoria; (b) títulos e ordem hierárquica de cada uma das seções do artigo; (c) parágrafos das seções e do Abstract; (c) indentações ou fórmulas associadas a sentenças finalizadas por “:.”; (d) referências e definições de notas de rodapé; (e) textos descritivos das figuras e tabelas, cujos corpos foram localizados e descartados; e (f) indexadores, autores, origem e ano das referências bibliográficas. O resultado deste trabalho é de grande utilidade para pesquisas baseadas em artigos técnicos, para os quais existe grande disponibilidade na internet. Este resultado preenche uma grande lacuna, deixada por detectores de sentenças que trabalham baseados somente na pontuação e sinalização de quebras de palavras na extração das sentenças, mas não são capazes de extrair sentenças misturadas aos vários tipos de informação que ocorrem quando um artigo técnico é convertido de formato PDF para o formato TXT.

Palavra-chave: processamento de linguagem natural, automação de ontologias, detecção de sentenças.