View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Servicio de Difusión de la Creación Intelectual

Uma Experiência de Utilização da Análise Semântica Latente Para o Tratamento de Documentos

Chu Chia Gean *Celso A. A. Kaestner*

Pontifícia Universidade Católica do Paraná (PUCPR)
Programa de Pós-Graduação em Informática Aplicada (PPGIA)
Rua Imaculada Conceição, 1155 – CEP 80.215-901
Curitiba – Paraná – BRASIL
{ccg, kaestner}@ppgia.pucpr.br

Resumo. Este artigo relata experimentos realizados para a realização automática de tarefas em Recuperação de Informações: recuperação e agrupamento de documentos. Nesta abordagem é empregada a Análise Semântica Latente (*Latent Semantic Analysis* - LSA), que emprega um método para a extração e representação da semântica contextual das palavras por meio de computações estatísticas aplicadas em uma coleção de documentos. A técnica LSA tenta explorar as relações semânticas "latentes" ou "implícitas" no texto, que são dadas pelas relações entre os termos, ao invés de considerar a semântica das palavras isoladas. Uma forma corrente de aplicar a LSA utiliza a técnica de decomposição em valores singulares (*Singular Value Decomposition* - SVD), como forma de redução da dimensionalidade do espaço de termos. A técnica empregada e sua aplicação à tarefas de recuperação e agrupamento são descritas por meio de sua aplicação a base de documentos padrão TREC, e os resultados obtidos são detalhados.

Abstract. In this paper we present experiments for the automatic execution of some important Information Retrieval tasks: the retrieval and the clustering of documents. In our proposal we employ *Latent Semantic Analysis* (LSA), which is a method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. The LSA technique tries to explore the "latent" or "implicit" semantics in the text, given by global relations among the terms, rather than use the semantics of the isolated words. One possible way to obtain LSA is the use of the Singular Value Decomposition (SVD) technique, in order to reduce the dimensionality of the working space. The technique is described over some representative applications in the standard TREC document-base, and the obtained results are presented.

Palavras-chave: Recuperação de Informações, Análise Semântica Latente, Decomposição em Valores Singulares

Keywords: Information Retrieval, Latent Semantic Analysis, Singular Value Decomposition

Submissão ao "IX Congresso Argentino de Ciencias de la Computación"