Similaridade entre Atributos de Entidades Baseada em Dados Obtidos da Web

PRISCILA SAD DE SOUSA (Autor), Anderson Almeida Ferreira (Orientador)

A similaridade desempenha um papel importante em diversas tarefas, tais como, Recuperação da Informação, Processamento de Linguagem Natural, Resolução de Entidades, dentre outras. Identificar a similaridade entre dois textos de atributos de duas entidades, como os títulos de entidades artigos científicos, pode não ser viável a partir da comparação direta desses textos, ou usando abordagens de similaridade baseada em conhecimento. A web, sendo uma fonte imensurável de informações, na qual é possível encontrar dados sobre quaisquer assuntos, pode auxiliar na verificação da similaridade entre esses pares. Neste trabalho, está sendo proposto um método dividido em etapas, onde, primeiramente, N páginas web são recuperadas, usando um motor de busca, para cada texto da comparação, em seguida é realizada a extração de dados dessas páginas web e posteriormente o pré-processamento. Logo depois, é feita uma avaliação usando lista de termos de gênero, segundo os princípios de coleta temática, para verificar se a página é relevante ou não ao contexto/tema de interesse. Caso seja relevante, é feita uma seleção dos termos próximos aos termos de gênero. O processo é realizado para selecionar as M páginas web mais relevantes dentre as páginas recuperadas anteriormente. A partir desses dados extraídos são geradas características usando modelo espaço vetorial com a ponderação dos termos usando TF-IDF e modelagem de tópicos. Uma função de similaridade é então proposta a partir da soma ponderada dessas duas características, cujos pesos e limiar foram definidos a partir da implementação da metaheurística GVNS. O trabalho encontra-se em andamento e ainda não foi comparado com outros métodos, no entanto, os resultados preliminares mostram que a função de similaridade tem apresentado resultados satisfatórios para auxiliar na identificação da similaridade entre um par de textos/valores de atributos de uma entidade.

Instituição de Ensino: Universidade Federal de Ouro Preto

ISSN: 0