

Determinação automática de limites de similaridade em processos de coleta temática de páginas da Web baseada em gênero

GUSTAVO OLIVEIRA DE SIQUEIRA (Autor), GUILHERME TAVARES DE ASSIS (DECOM) (Orientador)

Coletores temáticos apresentam o propósito de coletar páginas da Web que sejam relevantes a um tópico de interesse específico do usuário. Visando melhorar a eficácia e a eficiência de processos de coleta temática, foi proposta e desenvolvida, pelo orientador deste projeto, uma abordagem para coleta temática onde o tópico de interesse pode ser expresso por termos que descrevem o conteúdo e o gênero (estilo) das páginas da Web desejadas. Tal abordagem possibilita a construção de coletores temáticos eficazes, eficientes e escaláveis, uma vez que os limites de similaridade definidos e utilizados em tais processos de coleta sejam adequados. O limite de similaridade é utilizado pela abordagem, em um processo de coleta, para verificar se uma página da Web visitada é ou não relevante em relação ao tópico de interesse desejado. Logo, para validar tal abordagem proposta para coleta temática baseada em gênero, foi estabelecido um limite de similaridade distinto, empiricamente, para cada tópico de interesse considerado. Neste contexto, no intuito de tornar a abordagem menos dependente do usuário, o objetivo principal desse projeto de iniciação científica consistiu na proposta e validação de estratégias para determinação automática de limites de similaridade a serem usados em processos de coleta temática da abordagem baseada em gênero. Tais estratégias visam determinar os valores de limites de similaridade por meio: (1) da média aritmética ou ponderada das similaridades das páginas-semente do processo de coleta, geradas automaticamente pela abordagem; (2) do uso dos métodos de agrupamento K-Means e BIRCH; ou (3) da maximização da métrica Coeficiente de Silhueta para qualidade de agrupamentos. Experimentos de validação das estratégias foram realizados, gerando, como melhor resultado, uma precisão de 99% e um F1 de 93% em um determinado processo de coleta que considerou o limite de similaridade obtido automaticamente: resultado bem satisfatório em relação ao baseline utilizado.

Instituição de Ensino: Universidade Federal de Ouro Preto