Desenvolvimento e validação de uma ferramenta distribuída para coleta temática de páginas da Web baseada em gênero

MARCOS VINICIUS OLIVEIRA SOUZA (Autor), GUILHERME TAVARES DE ASSIS (DECOM) (Orientador)

Coletores temáticos apresentam o propósito maior de coletar páginas da Web que sejam relevantes a um tópico ou interesse específico do usuário, sendo importantes para uma grande variedade de aplicações. Em geral, eles funcionam tentando localizar e coletar páginas que estejam relacionadas a um determinado tópico de interesse. 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 desejado pode ser expresso por termos que descrevem o conteúdo e o gênero das páginas da Web desejadas. Tal abordagem baseada em gênero possibilita a construção de coletores temáticos que realizam processos de coleta eficazes e eficientes, conforme demonstrado experimentalmente. Nesse contexto, com o objetivo de melhorar a escalabilidade da abordagem para coleta temática baseada em gênero, este projeto de iniciação científica propôs uma arquitetura nova de funcionamento para tal abordagem, onde etapas relativas a processos de coleta temática podem ser realizadas de forma distribuída. Para validação dessa arquitetura, foram realizados experimentos iniciais que consistiram em processos de coleta de páginas da Web, relativas a um tópico de interesse, considerando a arquitetura distribuída proposta e a forma original de funcionamento da abordagem para coleta temática (baseline) sem envolver processamento distribuído. Para ambos processos, foi medido o tempo de execução; para tanto, como critério de parada dos processos, foi considerada a visita de 100.000 páginas da Web. Os experimentos realizados comprovaram a melhoria da escalabilidade da abordagem em relação ao baseline: de uma forma geral, considerando 8 computadores, o ganho foi de, aproximadamente, 83% considerando o tempo de execução total dos processos de coleta temática.

Instituição de Ensino: Universidade Federal de Ouro Preto

ISSN: 21763410