Ciência e Tecnologia no caminho
da Cooperação Internacional

Mineração de dados textuais eletrônicos - um software para reconfiguração de arquivos XML

Hélvia Pereira Pinto Bastos, Breno Fabrício T. Azevedo,
Yasmmin Côrtes Martins

IV Congresso Fluminense de Iniciação Científica e Tecnológica

17º Encontro de IC da UENF
9º Circuito de IC da IFF
5ª Jornada de IC da UFF

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA FLUMINENSE

Ciência da Computação

RESUMO

Introdução Este projeto tem por objetivo desenvolver ferramentas para análise de conteúdo e relevância de textos redigidos em fóruns de discussão e chats. Nesta etapa do projeto foi produzido um software para reconfigurar os arquivos XML resultantes do processamento feito no software Analisador de Presença Social (Kambara-Silva, 2011). Objetivos Um dos objetivos do projeto foi desenvolver um software para reconfigurar o conteúdo dos documentos XML gerados por um analisador de presença social (Bastos, 2012). Após a formatação dos arquivos XML, eles são classificados e agrupados de forma correta pelo software de mineração e agrupamento denominado Eurekha (Wives, 2004). Metodologia 1. Estudo do software Eurekha. 2. Estudo do software Analisador de Presença Social. 3. Construção do software para reconfigurar os arquivos XML. Resultados O software para reconfigurar os arquivos XML foi desenvolvido na linguagem de programação Java. O procedimento utilizado para fazer o rearranjo das informações contidas no documento foi obtido pela análise dos resultados gerados pelo Eurekha, como a lista de palavras e seus valores de relevância. Os arquivos XML quando colocados no Eurekha, antes de passarem pela reconfiguração, eram agrupados de forma desorganizada e não continham uma lógica coerente de classificação. Os arquivos permaneciam concentrados, pois o algoritmo usado na classificação (best-star) criava a relevância a partir das tags que eram padrões para todos os arquivos. Os valores numéricos encontrados nos documentos eram interpretados como "palavras adicionais", fazendo-se uma interpretação e, portanto, uma classificação indesejada. Com a reconfiguração dos arquivos, as quantidades ao invés de serem transcritas em números, foram descritas repetindo o nome da tag a qual elas pertencem. Assim é possível estabelecer a relevância e a frequência de forma correta. Esta operação foi executada para cada arquivo XML. Observou-se que os arquivos ficaram mais distribuídos após a reconfiguração. Os documentos de cada grupo possuíram características semelhantes em relação às suas pistas textuais e quantidades. Conclusões A partir dos resultados, depreende-se que o software desenvolvido apresentou resultados mais compreensíveis e coerentes ao contexto do agrupamento desejado. Logo, ele pode ser

ISSN 2177-6180

utilizado como uma ferramenta de transição entre a análise feita pelo Analisador de presença social e sua interpretação e classificação no Eurekha.

PALAVRAS CHAVE: Agrupamento, Classificação, Analisador.

APOIO FINANCEIRO: CNPQ

IV Congresso Fluminense de Iniciação Científica e Tecnológica

17º Encontro de IC da UENF
9º Circuito de IC da IFF
5ª Jornada de IC da UFF

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA FLUMINENSE

Ciência da Computação

ISSN 2177-6180