

Título: Sistemas inteligentes para texto da WEB

Autor(es) Fernando Hideo Fukuda*

E-mail para contato: fernando.fukuda@estacio.br

IES: PUC-RJ

Palavra(s) Chave(s): Web Text Mining (Wtm); Knowledge Discovery In Text (Kdt); Artificial Neural Network (Ann); Data Mining (Dm)

RESUMO

Esta dissertação investigou a aplicação dos processos de Knowledge Discovery in Databases (KDD) e técnicas de inteligência artificial para a criação de uma ferramenta de avaliação de textos da WEB de acordo com um determinado perfil. O trabalho, na área de sistemas de computação, teve como objetivo o desenvolvimento de técnicas que em conjunto com técnicas de inteligência computacional podem automatizar a avaliação de textos. Foram ainda pesquisados e implementados algoritmos inéditos para extração de características e avaliação de textos. A validade destes algoritmos foram comprovadas através dos resultados dos estudos de casos apresentados neste trabalho. Uma outra contribuição deste trabalho é a pesquisa de técnicas de inteligência computacional baseadas em sistemas especialistas e rede neurais para a avaliação de textos sem a necessidade de um Processador de Linguagem Natural (PLN), tornando-as independentes da linguagem natural escrita utilizada nos textos dos documentos. Esta característica permite a utilização destas técnicas em documentos escritos em diferentes línguas, como foi comprovado neste trabalho, considerando-se textos em português e inglês. Além disso, foi desenvolvido e implementado um processo inédito de Knowledge Discovery in Text (KDT). Um outro objetivo concluído foi a implementação do sistema computacional SITEX - Sistema Inteligente para Textos da Web - desenvolvido em linguagem de programação C++, usando-se o processo de KDT, as técnicas de inteligência computacional e o algoritmo desenvolvido nesta pesquisa. Os estudos de casos foram realizados com textos fornecidos pela WE, através de inscrições em jornais on-line e avaliados de acordo com três tipos de perfis: economia, esportes e informática. Os estudos de casos comprovam que a rede neural não é dependente do perfil. Isto significa que os algoritmos descobertos nesta pesquisa possuem a capacidade de extrair conhecimento do perfil para a base de conhecimento e os valores calculados das características de representação dos textos que indicam se o texto pertence ou não ao perfil de interesse. Esta condição, oferece uma grande robustez ao sistema. A realização dos estudos de casos, com serviços e textos reais da Internet, comprova a utilidade destas técnicas em aplicações reais. A comparação dos resultados do SITEX superiores aos resultados do TextAnalyst, produto de mercado para análise de textos da Web baseado em inteligência computacional, comprovam a qualidade do processo, das técnicas e do algoritmo apresentados nesta dissertação. Os resultados comprovam a validade do processo, das técnicas e do algoritmo desenvolvidos e implementados neste trabalho: (i) acertos sobre arquivos dentro do perfil de 83,33%; (ii) acertos sobre arquivos fora do perfil de 96,36% e (iii) acertos sobre todos os arquivos de 93,15%. Estes resultados atendem a principal motivação deste trabalho que é a verificação do crescimento explosivo de documentos disponíveis a Web e da necessidade das pessoas utilizarem uma ferramenta de auxílio para o tratamento e seleção destes documentos de uma forma inteligente e personalizada. Todos os objetivos propostos neste trabalho foram concluídos com sucesso.