Proposta de projecto de colaboração com o Arquivo da Web Portuguesa

Reconhecimento de siglas em conteúdos arquivados da Web

A FCCN tem em curso o projecto de Arquivo da Web Portuguesa e procura colaborar com entidades de Investigação e Desenvolvimento, que tenham interesse em participar na realização de projectos inovadores. Este documento apresenta uma proposta de um projecto com a duração estimada de 1 ano, que poderia fazer parte de um trabalho de mestrado ou de iniciação à investigação.

Por vezes, as siglas existem apenas durante um período de tempo, deixando de ser referidas em documentos que sejam produzidos a partir dessa época. Isto provoca uma dificuldade acrescida para alguém que se depare com uma sigla que não seja actual ou cujo documento em que ela consta não contenha a sua explicação. Por outro lado, as siglas podem ter vários significados, que podem coexistir ou ir evoluindo ao longo do tempo (ex.: GNR – Guarda Nacional Republicana, Grupo Novo Rock; RTP – Radiotelevisão Portuguesa, Rádio e Televisão de Portugal).

À medida que as páginas vão deixando de existir na Web, visto que a volatilidade dos conteúdos na Web é alta, também os motores de busca vão deixando de apresentar resultados para siglas em desuso. Ao procurar o significado de uma sigla a partir de um motor de busca, poderá ser mais difícil encontrar resultados para siglas que já não se usem, embora tenham importância histórica uma vez que foram usadas em documentos no passado.

O objectivo deste projecto é criar um sistema automático de reconhecimento de siglas em conteúdos arquivados. O produto deste projecto será integrado no serviço público do Arquivo da Web Portuguesa, para que os utilizadores tenham uma funcionalidade adicional de pesquisa de siglas.

O sistema poderá ser implementado na linguagem JAVA sobre a tecnologia Hadoop, uma implementação open source do paradigma de programação MapReduce desenvolvido pelo Google. Esta tecnologia permite distribuir e paralelizar processamentos por clusters com milhares de processadores, sobre quantidades de dados na ordem de grandeza dos Petabytes. Esta escalabilidade quase ímpar e atingida com reduzido esforço para o programador, está

actualmente a ser aproveitada por o Yahoo em mais de 10.000 máquinas, para processar até 1 Petabyte de dados em diversos estudos e tarefas, inclusive na indexação de toda a Web para o seu motor de busca.

Bibliografia

- Acronym Finder (http://www.acronymfinder.com/)
- Dannélls, D. 2006. Automatic acronym recognition. In *Proceedings of the Eleventh Conference of the European Chapter of the Association For Computational Linguistics: Posters & Demonstrations* (Trento, Italy, April 05 - 06, 2006). European Chapter Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 167-170.
- Larkey, L. S., Ogilvie, P., Price, M. A., and Tamilio, B. 2000. Acrophile: an automated acronym extractor and server. In *Proceedings of the Fifth ACM Conference on Digital Libraries* (San Antonio, Texas, United States, June 02 - 07, 2000). DL '00. ACM, New York, NY, 205-214. DOI= http://doi.acm.org/10.1145/336597.336664
- Sánchez, D. and Isern, D. 2009. Seeking Acronym Definitions: a Web-based Approach. In *Proceedings of the 12th international Conference of the Catalan Association For Artificial intelligence* S. Sandri, M. Sànchez-Marrè, and U. Cortés, Eds. Frontiers in Artificial Intelligence and Applications, vol. 202. IOS Press, Amsterdam, The Netherlands, 339-348.
- Stuart Yeates. 1999. *Automatic extraction of acronyms from text*. Proc. of the Third New Zealand Computer Science Research Students' Conference. University of Waikato, New Zealand.
- Torii, M., Liu, H., Hu, Z., and Wu, C. 2006. A comparison study of biomedical short form definition detection algorithms. In *Proceedings of the 1st international Workshop on Text Mining in Bioinformatics* (Arlington, Virginia, USA, November 10 - 10, 2006). TMBIO '06. ACM, New York, NY, 52-59. DOI= http://doi.acm.org/10.1145/1183535.1183548
- Xu, J. and Huang, Y. 2006. Using SVM to Extract Acronyms from Text. *Soft Comput.* 11, 4 (Nov. 2006), 369-373. DOI= http://dx.doi.org/10.1007/s00500-006-0091-5
- Yeates, S., Bainbridge, D., and Witten, I. H. 2000. Using Compression to Identify Acronyms in Text. In *Proceedings of the Conference on Data Compression* (March 28 - 30, 2000). DCC. IEEE Computer Society, Washington, DC, 582.
- Youngja Park and Roy J. Byrd. 2001. Hybrid text mining for finding abbreviations and their definitions. *In Proceedings of EMNLP 2001*, pages 126--133.