Proposal for a collaborative project with the Portuguese Web Archive

Recognition of abbreviations and acronyms in archived Web content

FCCN is currently engaged in the Portuguese Web Archive project and seeks to cooperate with Research and Development organisations who are interested in participating in innovative activities. This document presents a proposal for a project with an estimated duration of 1 year, which could form part of a master's thesis or introduction to research.

Sometimes abbreviations exist for a limited period of time and then become obsolete. This causes difficulty for someone faced with an abbreviation or acronym which is not currently used and is not explained the text where it is used. In addition, an abbreviation can have multiple meanings that can co-exist or evolve over time (e.g. NASA: National Aeronautics and Space Adminsitration, North Andover Soccer Association, Native American Student Association; CIA: Central Intelligence Agency, Cleveland Institute of Art).

Because the volatility of Web content is high, as pages are removed from the Web, search engines also stop delivering results for abbreviations which have fallen into disuse. When trying to find the meaning of an abbreviation from a search engine, it may be more difficult to find results for abbreviations that are no longer in use, even though they might have historical importance given their use in past documents.

The aim of this project is to create a system for the automatic recognition of abbreviations and acronyms in archived content. The product of this project will be integrated into the public service of the Portuguese Web Archive, so that users also have an abbreviation search feature.

The system may be implemented using JAVA with Hadoop technology, an open-source implementation of the MapReduce programming paradigm developed by Google. This technology allows distributed and parallel processing on clusters with thousands of processors. This almost unmatched scalability, achieved without too much effort from the programmer, is currently being used by Yahoo! on more than 10,000 machines to process up to one Petabyte of data

in several studies and tasks, including indexing the entire web for its search engine.

Bibliography

- 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.