Proposal for a collaboration project with the Portuguese Web Archive

Automatic classification of archived web content

The Portuguese Foundation for National Scientific Computing (FCCN) is working on the Portuguese Web Archive and aims to co-operate with Research and Development entities interested in participating in innovative projects. This document presents a proposal for a project with an estimated duration of 1 year.

The Portuguese web is regularly collected and stored for future preservation. This large amount of data requires mechanisms that enable information to be efficiently searched and accessed. Frequently, the search space to be refined to identify relevant results.

Document classification helps to respond to these necessities, for instance, making it possible to browse hierarchically through class trees where the documents are grouped according to topic. The Yahoo and Dmoz search directories are examples of this paradigm. However, they require a strong human intervention to categorize information which cannot be applied in a web archive due to the large amount of historical data involved.

The objective of this project is to create an automatic classification system for web documents archived over time in the Portuguese Web Archive to support faceted search or clustering results. Classification would be performed at two levels, by topic or sub-topic, identifying the main topics addressed in the documents (e.g. sports→football, politics→international).

This way, it would be possible to look for and extract all documents regarding a given topic. This classification would enable search results by term to be grouped in the Web Archive, with two objectives: the first, as a visual clue to refine searches and the second, to increase the variety of results by topic**.**

A specific aspect of a web archive is that the different archived versions of a page might have evolved across time, both from a visual and content point of view, therefore the addressed topics might also have evolved.

This project must address significant challenges. For instance, although most of the archived pages are written in Portuguese, there is a significant amount of archived content written in other languages and adequate training data sets are necessary to tune the classification software. Additionally, the classification algorithms must scale to process millions of documents in relatively short intervals of time.

The classification software should be implemented using the JAVA language. The objective is to run software over the archived data using Hadoop technology, an open-code implementation of the programming paradigm MapReduce developed by Google. This scalability is reached with reduced effort for the programmer and is currently being used by Yahoo on more than 10,000 machines, for various studies and tasks, even for indexing the entire web for its search engine.

Bibliography

- Soumen Chakrabarti. 2003. Mining the Web (Part II - Learning - supervised learning) .
- Michelangelo Ceci, Donato Malerba. Classifying web documents in a hierarchy of categories: a comprehensive study.
- Dumais, S. and Chen, H. 2000. Hierarchical classification of Web content.
- Xing, D., Xue, G., Yang, Q., and Yu, Y. 2008. Deep classifier: automatically categorizing search results into large-scale hierarchies.
- Cai, L. and Hofmann, T. 2004. Hierarchical document categorization with support vector machines.
- Rousu, J., Saunders, C., Szedmak, S., and Shawe-Taylor, J. 2005. Learning hierarchical multi-category text classification models.
- Kules, B., Kustanowitz, J., and Shneiderman, B. 2006. Categorizing web search results into meaningful and stable categories using fast-feature techniques.
- Zhang, D. and Lee, W. S. 2004. Web taxonomy integration using support vector machines.
- Hao, P., Chiang, J., and Tu, Y. 2007. Hierarchically SVM classification based on support vector clustering method and its application to document categorization.

Related Software

- http://lucene.apache.org/mahout/
- http://cwiki.apache.org/MAHOUT/
- http://mallet.cs.umass.edu/classification.php
- http://alias-i.com/lingpipe/demos/tutorial/classify/read-me.html
- http://www.cs.waikato.ac.nz/ml/weka/
- http://weka.wiki.sourceforge.net/Text+categorization+with+Weka

Training data sets

- http://rdf.dmoz.org/