Proposal for a collaborative project with the Portuguese Web Archive

Named Entity recognition in content of web archive

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.

The Portuguese web is periodically crawled and archived for future preservation. This large amount of data requires mechanisms that enable access to archived information. Therefore, search refinements must be supported to identify relevant documents to the performed queries.

The aim of this project is to create an automatic system to generate additional meta-data that describes the archived content. This will be based on the identification of named entities (NE). That is, it will locate and identify elements in the text associated with predefined classes, such as person or organization (e.g. Al Gore, United Nations) or temporal expressions (e.g. 6 August 1997).

Named entities recognition identifies particularly relevant information on web pages, which contributes to filter out noise and reduce indexes size. It also makes it possible to disambiguate searches performed by users, for example, differentiating JAVA the programming language from Java the Indonesian island, and to understand that certain sets of words have their own meaning (e.g. *President of the United Nations*).

In a web archive it is essential to find stored information according to its place in time. Current search mechanisms use the date of crawl to locate a page in time. However, this approach is vulnerable because a page may have been published on a previous date. The identification of temporal NEs will enable dates to be assigned to content more accurately and, as a result, will improve the quality of search results.

Named entity recognition software, such as Rembrandt, may be applied in this project. The named entity recognition software should be implemented preferentially in JAVA to be run on Hadoop technology, an open-source implementation of the MapReduce programming paradigm developed by Google that enables distributed and parallel processing on clusters composed by thousands of processors. This almost unmatched scalability is achieved with limited effort from the application programmer.