Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 385-386 recibido 1-05-2009; aceptado 5-06-2009

# QEAVis: Quantitative Evaluation of Academic Websites Visibility[1]

## *QEAVis: Evaluación Cuantitativa de la Visibilidad de los Sitios Web Académicos*[2]

**M. F. Verdejo, E. Amigó, L. Araujo, V. Fresno, G. Garrido, R. Martínez, A. Peñas, A. Pérez, J.R. Pérez, A. Rodrigo, J. Romo, A. Zubiaga**
NLP&IR Group, UNED
Dpto. Lenguajes y Sistemas Informáticos
C/ Juan del Rosal nº 16, 28040 Madrid

**I. Aguillo, M. Fernández, A. M. Utrilla**
Cybermetrics Lab
CCHS - CSIC
Albasanz, 26-28, 3C1.
28037 Madrid

**Resumen:** El proyecto plantea la aplicación de las TLH a un problema importante como es medir la visibilidad académica en la web, sentando las bases de una evaluación cuantitativa del compromiso de los departamentos universitarios con la accesibilidad pública de su información. Para ello es necesario desarrollar indicadores web (Cibermetría) y estudiar la visibilidad de los sitios web académicos, haciendo especial énfasis en la presencia del español (de importancia estratégica) y en el ámbito de las disciplinas relacionadas con humanidades (que requiere una ayuda especial respecto a su posicionamiento en web).
**Palabras clave:** Cibermetría, Categorización, Extracción de Información

**Abstract:** The project proposes the application of HLT to an important problem such as the measurement of the academic visibility in the web, giving the basis of a quantitative evaluation of the universities departments' commitment in the public access to their information. Web indicators (Cybermetrics) must be developed and applied to the study of the academic websites visibility, with special focus on the presence of the Spanish language (of strategic importance) and the academic areas related to humanities (which need special help for their web positioning)
**Keywords:** Cybermetrics, Categorization, Information Extraction

## *1 Objectives*

The application of Human-Language Technologies (HLT) to mine the web in order to automatically measure the academic visibility poses new technological challenges. The main goals of the QEAVIS project are the following:

1. To advance the state of the art of classification and extraction techniques for automating effectively the process of identifying and obtaining relevant information from websites. In particular the data needed to evaluate the impact of that website in the context of a research community
2. To advance the state of the art on web indicators and the methodological approach to test their reliability.
3. To gain insight on the presence and impact of the Spanish Humanities fields in the WWW

## *2 Groups involved and approach*

The project is organized in two subprojects, to be carried out by *NLP&IR-UNED* and the *Cybermetrics Lab* (CCHS-CSIC) respectively. UNED expertise includes crawling, multilingual information retrieval, classification and extraction techniques. The Cybermetrics Lab has developed cybermetrics methods to analyse and rank (mainly in an intellectual way) web sites visibility.

[1] Financiado por el Ministerio de Ciencia e Innovación TIN2007-67581

[2] http://nlp.uned.es/qeavis/

ISSN: 1135-5948 

In QEAVIS we will determine first the main web sites of academic contents at web sub domain level. These subdomains are crawled to download, store and manage their web pages, so that the web pages are prepared for their automatic classification and information extraction. Web subdomains will be classified under language, academic category and discipline. Furthermore, the information necessary for creating the microformat of each subdomain should be automatically extracted. This information will be used to elaborate a profile and a description of each university department. Finally, a variety of web indicators will be applied to the information of the subdomains in order to quantify their presence, visibility, impact and popularity. The resultant quantitative values will be used to make a ranking of subdomains/departments per each academic category. In the ranking, the top positions will be for those departments whose commitment to the visibility of their information is the largest.

The rankings, together with the criteria used in their construction, the recommendations and resources in order to improve the results, will be made public. So, we expect : (1) to stimulate the continuous improvement of the accessibility and visibility of the academic information in the web; (2) to provide new cybermetric indicators with finer granularity.

## *3 Current State of the project*

The Cybermetrics Lab has developed an automatic method for extracting sub domains of academic websites using the capabilities of the Yahoo! search engine. For a large population of the world top universities a large list was compiled. In order to select only those websites referred to research groups and departments still operative, a cleaning process was done with the help of automatic link checkers and manual editing. This is relevant for the standardisation of the name according to the Title tag's information. The resulting list has been the target for the automatic classification tools and for testing their precision.

A previous database consisting mainly of departments of European Universities was available with UNESCO codes. This database was classified manually a few years ago and it served as a testbed for fine tuning some of the tools. UNED crawled around 100,000 web sub domains at depth 3, with around 7,000,000 pages and documents. In a next step the indexing of this collection and its processing for obtaining the web and hosts graphs will be tackled. These graphs together with the statistics about rich documents, inlinks, outlinks, etc. will be the input for developing new cybermetric indicators.

Concerning terminology extraction, UNED has developed a tool to extract automatically relevant terminology from the term index corresponding to a web domain. This has been done by calculating the divergence between language models. Then, UNESCO codes have been used to select those pages from the collection which correspond to a particular academic domain in different levels of detail. Furthermore an index for each level using the documents of the category associated to the level has been generated. Applying the tool for terminology extraction to the index, term lists have been generated, characterizing each considered academic domain. These lists are used in the classification phase.

With regard to automatic classification, UNED has carried out a study on whether unlabeled data could improve results for multiclass web page classification tasks using Support Vector Machines (SVM). In the light of the results, it was decided to rely only on labelled data, both for good performance and for reducing the computational cost. Next step aimed at reducing the number of UNESCO categories to the 20-25 that better fitted our crawled collections, without taking into account the deep level of the taxonomy. To evaluate the adaptation of the collections to that taxonomy, an unsupervised grouping was tried: a clustering of the UNESCO categories represented by means of all the documents belonging to them was carried out using a Self-Organizing Map (SOM). The inconsistent results led to dismiss the use of UNESCO taxonomy in the future.

Based on the conclusions of the clustering results, the Cybermetric group decided to create a new taxonomy according to the domain we deal with: a combined multilingual (UDC, LC, UNESCO), with a hierarchical association of over 500 entries grouped in about 26 large categories.

With this new taxonomy and a training collection of manual categorized web pages, UNED has developed a supervised SVM classifier which is now in an evaluation phase.

Papers describing in detail the work done are available at the project website.