Organization of Web Document Collections Based on Link Semantics

Objectives

The requirements for effective search and management of the WWW are stronger than ever. THESUS is a system that aims at characterizing, organizing and querying large collections of Web pages.

Our main effort in the characterization of pages is using the incoming links of documents on the Web and extracting their semantics.

Innovation & Strong Points

- A model and language that enable thematic selection of WWW subsets, and their subsequent enrichment.
- A mechanism that extracts keywords and enhances documents' hyperlinks with semantics, by mapping sets of keywords that describe a web page to sets of concepts organized in a hierarchy.
- A novel similarity measure between web documents, that takes into account the similarity between weighted words in a hierarchy.
- Organising documents into sematic clusters, and labeling these clusters.
- The implementation of a client/server system, called THESUS.

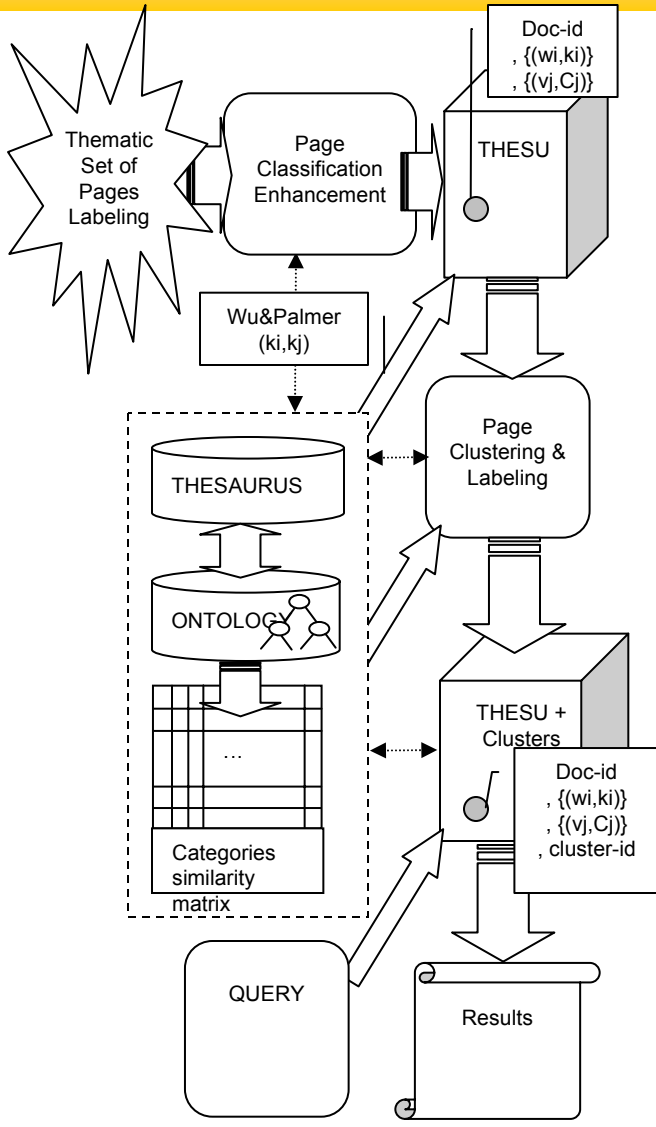
Impact

- A better characterization of documents on the web.
- An easier way to construct and organise thematic warehouses, with data from the web in a semi-automatic process.
- Querying documents not only with exact matches, but with similarity queries.

Partners

- Athens University of Economics and Business (AUEB) Computer Science Department, DB-NET.

Architecture



Results

- Querying links semantics yields correct information on the semantics of pages, at least comparable to full text querying.
- The clustering algorithm, based on the similiraty measure between pages returns meaningful clusters.
- Incoming link semantics can help establish a ranking of important thematic hubs on the Web.
- The whole system is a powerful tool that facilitates the construction of thematic warehouses using data from the Web.

Contact projet : NGUYEN Benjamin, INRIA projet GEMO
Domaine de Voluceau – Rocquencourt BP105, 78153 Le Chesnay Cedex France
01 39 63 51 88 ou Benjamin.Nguyen@inria.fr