

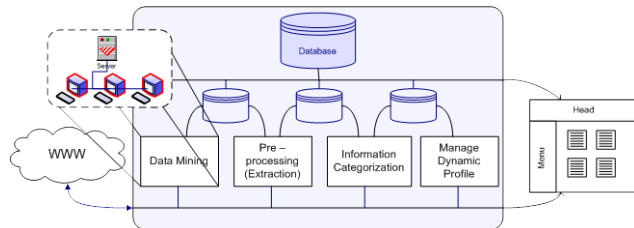
Research Activities – Web Technologies

Web Technologies include procedures that are used in order to enhance the services that are offered by the World Wide Web. They include both services that can be presented directly to the users of the World Wide Web and services that are transparent to the end user but help in order to promote the usability of the Web.

Web Technologies can be separated to many categories. Research Unit 6 focalizes on Dynamic Processing of Web Content (Information Extraction), Data Mining, Web Content Summarization, Categorization and Personalization, and Web Site Construction.

The basis of most of the aforementioned Web Technologies is collecting data and extracting keywords. This is a procedure that includes algorithms that are close to lexicographic analysis of the content. Although this analysis could not be thought as a web technology, it still remains an essential part. Moreover, some algorithms for web technologies require the existence of metadata in order to acquire information about the meaning of the content as lot of information include more than text (multimedia).

The languages that are used in our research are basically C++, Java, XML, HTML, PHP, ASP.NET, JSP, MSSQL, MySQL and more.

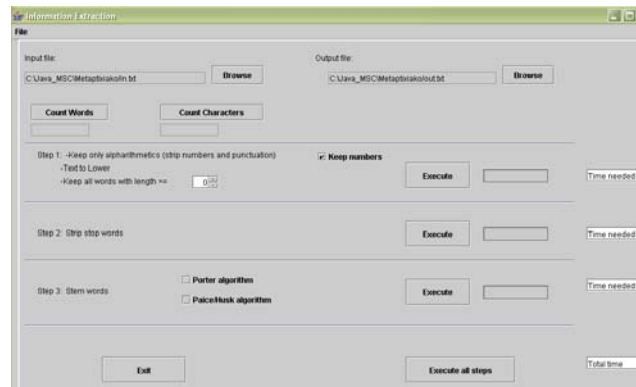


Dynamic Processing of Web Content – Information Extraction

In the chaotic conditions that occur in the web content with the mass creation of web data, Web Technologies are becoming a trend – more than a tool – in order to help the users of the Internet to their navigation. More specifically, dynamic processing of Web Content is a technology that leads to Information Extraction.

Information Extraction is the procedure of extracting the meaning of a text provided that specific type of metadata and simple data are given to the Information Extraction Mechanism.

The basis of Information Extraction is text analysis and more accurately keyword recognition and extraction. Research Unit 6 has put the focus on services that could be used for keyword extraction and thus we are experimenting on creating Information Extraction Systems.



Web Content Summarization and Categorization

Another crucial part of Web Technologies is further processing of the information that is extracted from web data. This further processing introduces algorithms, procedures and tools that lead to summarization of large amount of data and automatic categorization. These technologies can be used by mechanisms that present data to the users of the web. Additionally, as technology expands, a need for fewer amounts of data arises especially for the users of small screen devices.

Web Content Summarization can be based on simple keyword extraction or on Information Extraction. In the first case, Research Unit 6, focuses on locating algorithms that could be both quick and effective in order to produce summaries for web content. Some simple but helpful examples of summarization could be web site summarization, article summarization, construction of tools for small screen devices and more.

Web Content Automatic Categorization is the basis of the new trend of the World Wide Web which is Web Personalization. Some Personalization Techniques are based on the actions of the users when using a device, when visiting a device, or when using a program that utilizes web content. The categorization consists of algorithms and procedures that, based on predefined categories, can detect the similarity of a text with a category and rank it according to the most similar category. Additionally, this similarity issue can be used in order to locate the matching between two texts. Our research analysis is focused on locating algorithms that could use a combination of SVM algorithm and positive and negative terms for each category in order to categorize web content.

Our intention is to create mechanisms for summarization and categorization, which can be used as modules to software, in order to enhance the usability and efficiency.

Web Site Construction - Personalization

All the aforementioned technologies introduce techniques that are transparent to the user because this is their intention and nature. What is tangible to the World Wide Web users are web sites and services that are offered through them. Research Unit 6 is doing research on web site construction of dynamic content and thus on technologies that include dynamic web languages. The research is put on constructing web sites focalizing on efficiency and re-usability without excluding the human factor and the graphic user interface.

Research Unit 6 has put the focus on Web Site Personalization which seems to be an essential part of all the dynamic content web sites. The Personalization issue that is implemented by Research Unit 6 is based both on the users' actions and on the combination of all the aforementioned technologies. We believe that this combination can lead to better results for enhancing the usability and results that derive for the construction and use of the personalized portals.

Recently, we have expanded our research of Web Technologies on web services, agents and focused crawlers in order to extend our systems with features that are introduced lately to web technologies and can enhance even more any of our modules, especially the services that are provided and promoted through the construction of dynamic web sites and portals.

The screenshot shows the CARAMELLA news portal. At the top, the word "CARAMELLA" is displayed in large, stylized letters, with "CATEGORIZE AND RAISE NEWS LESS LABOR" written below it. A navigation bar includes links for "Central Page", "Register", "Contribute", "Contact", and "Info". On the left, there is a "Login" section with fields for "UserName:" and "PassWord:". Below this is a "Category" list with links for "business", "education", "entertainment", "health", "politics", "sports", and "technology". The main content area is divided into two columns. The left column, titled "CARAMELLA", contains a welcome message and a description of the portal's technology. The right column, titled "HOT NEWS", features a headline: "A radical restructuring plan for the Italian airline Alitalia has been approved by the European Commission. Rival European airlines had opposed the plan, arguing it included illegal state aid for the loss-making carrier..."

CONTACT INFORMATION

RESEARCH UNIT 6

University Campus, Building B', GR-26500 Patras, Greece

Tel.: (+30) 2610 960375, 2610 960355

2610 960380, 2610 960316

Fax: (+30) 2610 960358

URL: <http://ru6.cti.gr>

E-mail: rd-unit-6@cti.gr

CONTACT PERSON

Dr. Christos Bouras (Professor)

Tel: (+30) 2610 960375

Fax: (+30) 2610 969016

URL: <http://ru6.cti.gr/bouras>

E-mail: bouras@cti.gr

COMPUTER TECHNOLOGY INSTITUTE & PRESS “DIOPHANTUS”

RESEARCH UNIT 6

NETWORKS TELEMATICS AND NEW SERVICES

Research Field:

Web Technologies



COMPUTER TECHNOLOGY INSTITUTE & PRESS “DIOPHANTUS”

