Bachelorthesis-Aufgabe

Web-document Acquisition System

ID	IBEO1-1-10
Studierende	Thomas Schumacher Micha Wyttenbach
Betreuer	Dr. Olivier Biberstein
Experten	
Aufgabe	This bachelor thesis aims at developing a web-document acquisition system to collect a huge amount of documents for further static analysis (statistics, data-mining, etc.). Different types of web-documents can be envisaged such as web-sites, emails, auction ads, etc.. Each type of documents requires a different acquisition mechanism. This thesis concerns the acquisition of web-sites only and an acquisition mechanism based on the web-crawling technique. However, the system should be flexible enough to add new types of web-documents and acquisition mechanisms in the future. The acquisition system has to deal with web-documents of interest only, i.e. web-documents related to some specific topics. In other words a fast and simple filtering system has to be implemented to keep or ignore a newly acquired web-document. Moreover, web-documents have to be parsed in order to extract specific information for further analysis. Both the parsing and filtering system depend on some pieces of information that should be parameters the application but not hard-coded. *Main Objectives* • Study of availability/usability of open source web-crawlers • Development of an application with a graphical user interface in charge of * acquisition of web-sites by means of web-crawling technique for a given depth * fast and simple filtering system to keep the documents of interest * parsing of web-documents and extraction of specific information for further analysis * storage of extracted information and acquired web-documents with their timestamp *Optional Objectives* • Study of a more advanced filtering system. • Construction of the underlying graph induced by the hyperlinks of a web-document