



Learning Content Models for Semantic Search

Eran Peer, Hila Shalom

Advisors: Prof. Michael Elhadad, Dr. Menachem Adler

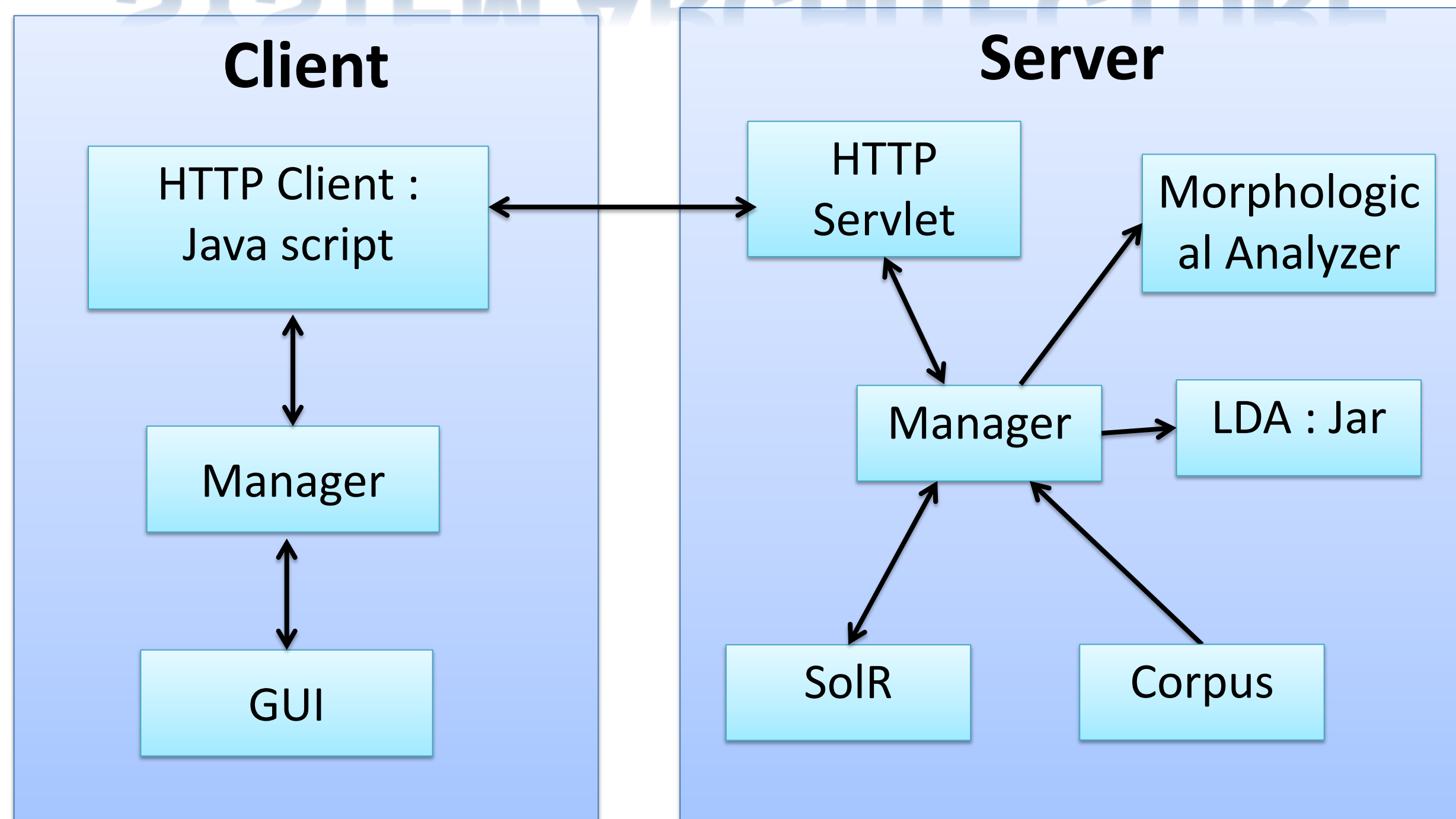
THE PROJECT GOAL

to allow the user to explore a repository of textual documents, and discover material of interest even when he is not familiar with the content of the repository.

Existing search engines provide powerful features to identify known documents by using a short description of their content (keywords, name of document). For example, a user wants to find information about a term that he knows, he enters the term's name in a search engine, and the search engine returns the address of the term's value at Wikipedia.

The task we address is what happens when the user does not know the name of the term he is looking for, or when the term the user enters has several meanings. The system that we developed will help the user explore interactively the repository and the terms that are most significant in this repository.

SYSTEM ARCHITECTURE



Client

- HTTP Client – The functions in the client side
- GUI – Web based graphical user interface

Server

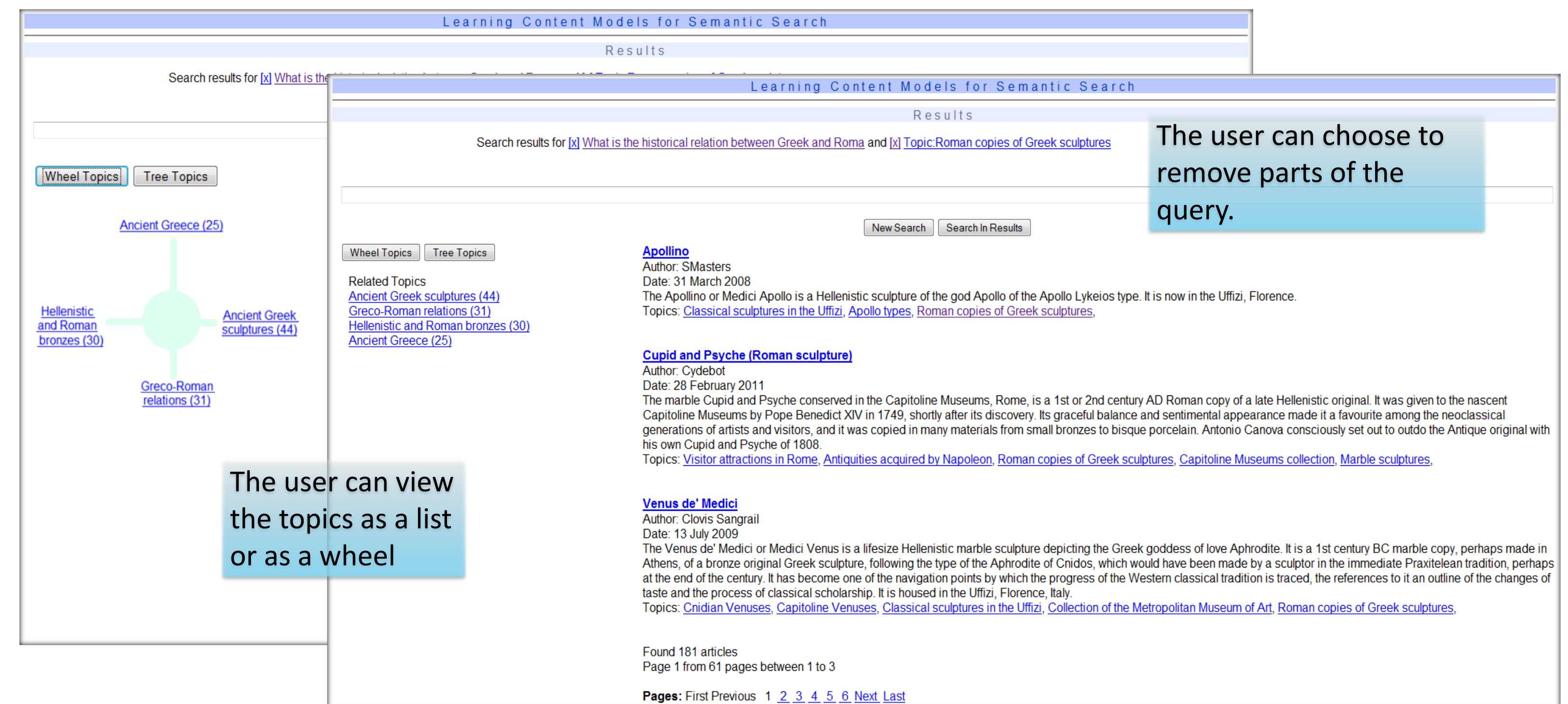
- HTTP Servlet – Handles the connection with the client
- Morphological Analyzer – Analyze the articles into verb, noun, adjective etc.
- LDA – Algorithm that creates hierarchical topics from the corpus. Tags all the articles with the topics. Can be replaced with a better one.
- Corpus – Contains the articles and their data (author, date etc.). New and existing articles can be added or removed. Can be at English or Hebrew.
- SolR – System for searching quickly and efficiently in huge corpus.

User Graphical Interface

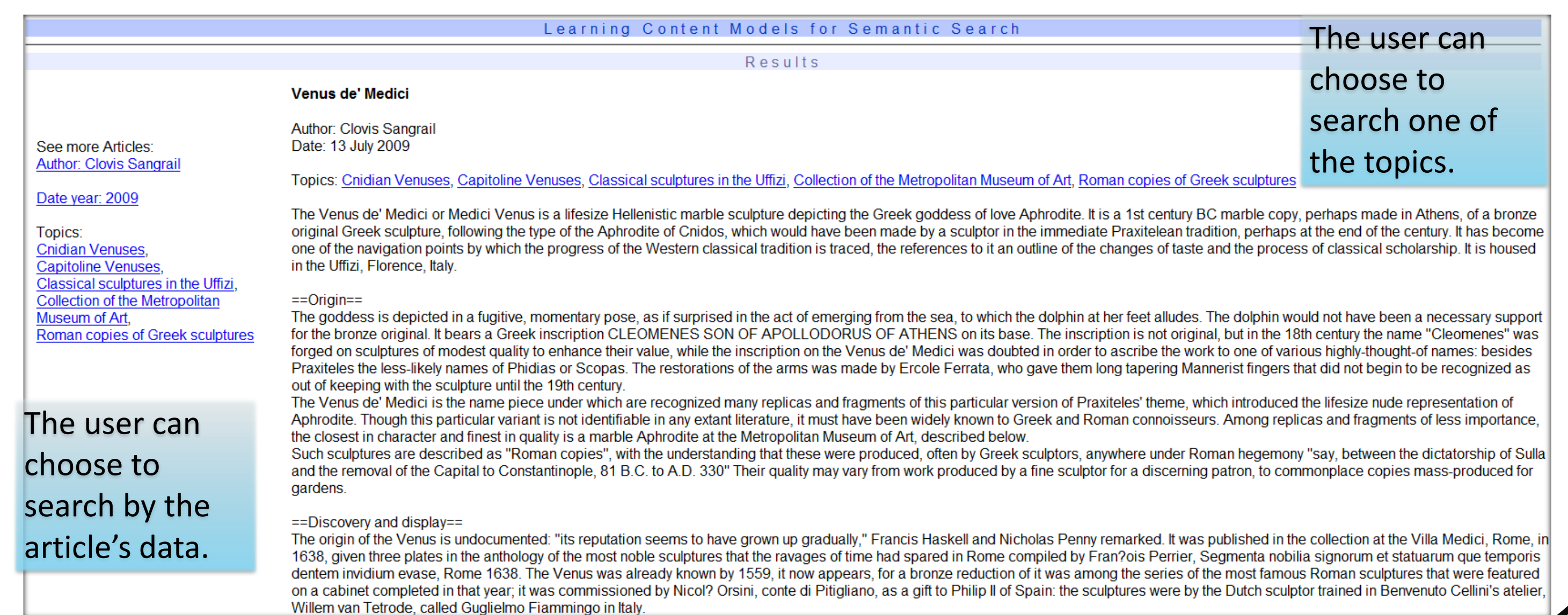
The user entered the query “What is the historical relation between Greek and Roma”. Here are the query’s results.



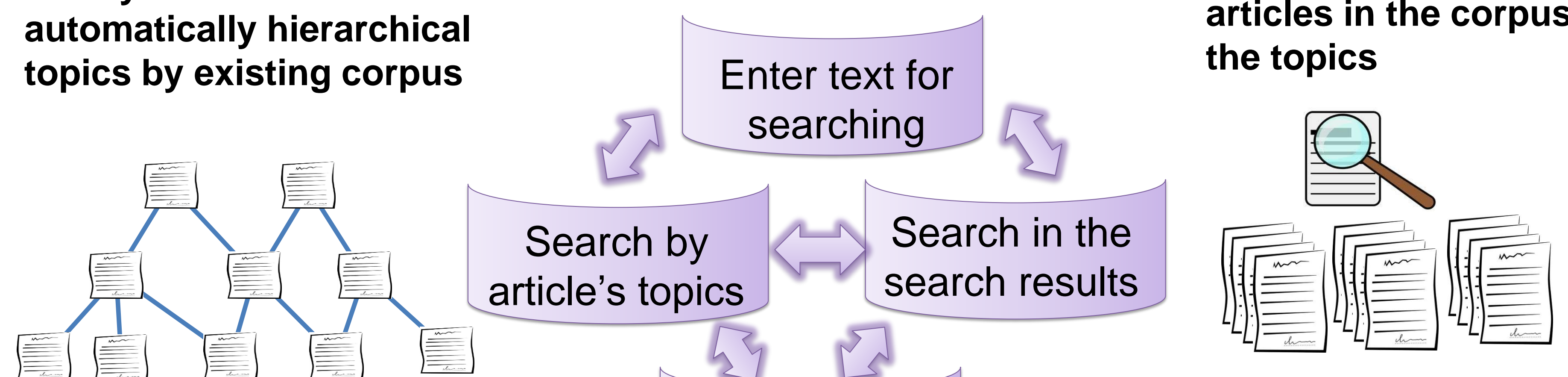
The user clicked the topic “Roman copies of Greek sculptures”. Here are the results to the query with the topic.



The user chose the article “Venus de’ Medici”. Here is the article.



The system builds automatically hierarchical topics by existing corpus



The users send questions in natural language (English/Hebrew), and receive relevant results



The users can navigate in the system to understand the corpus



The system tags the articles in the corpus with the topics



Web system - available every where, every time



This project was programmed in Java and Javascript using Eclipse.

The system was developed using Mallet, SolR, Morphological analyzer and Tomcat.



Our code at <http://code.google.com/p/semantic-search2011>