

SEE-GRID Regional Application

SE4SEE

A Grid-Enabled Search Engine for
South Eastern Europe



SEE-GRID

South Eastern European GRid-enabled
eInfrastructure Development

*B. Barla Cambazoglu, Ata Turk, Evren Karaca,
Cevdet Aykanat, Bora Ucar, and Tayfun Kucukyilmaz*
Bilkent University, Computer Engineering Department

Onur Temizsoylu
TÜBİTAK/ULAKBİM

What is SE4SEE?



SEE-GRID

South Eastern European GRid-enabled
eInfrastructure Development

SE4SEE is a

- personalized
- on-demand
- focused
- category-based
- grid-enabled

search engine for the countries in South East Europe.

Components of SE4SEE



SEE-GRID
South Eastern European GRid-enabled
eInfrastructure Development

SE4SEE is composed of three components

- a user interface (Web portal)
- a Web crawler
- a text classifier

Web Crawling Component



SEE-GRID
South Eastern European GRid-enabled
eInfrastructure Development

A web crawler (spider) is a program that

- locates,
- downloads, and
- stores web pages.

The Web crawler in SE4SEE is

- written in Java
- multi-threaded
- based on the Websphinx library

Grid-Enabled Web Crawling



SEE-GRID
South Eastern European GRid-enabled
eInfrastructure Development

The size of the Web is enormous; Web crawling requires:

- high download rates over the network,
- excessive storage (both memory and disk),
- vast amount of processing power.

It is hard to crawl the Web by a centralized system

Solution: Grid-enabled Web crawling since

- excessive disk and memory capacities together with the huge CPU power of the Grid provides the necessary medium for storing and processing the vast Web content,
- geographically distributed grid nodes allow fast download rates and dispersion of the network load.

Text Classification Component



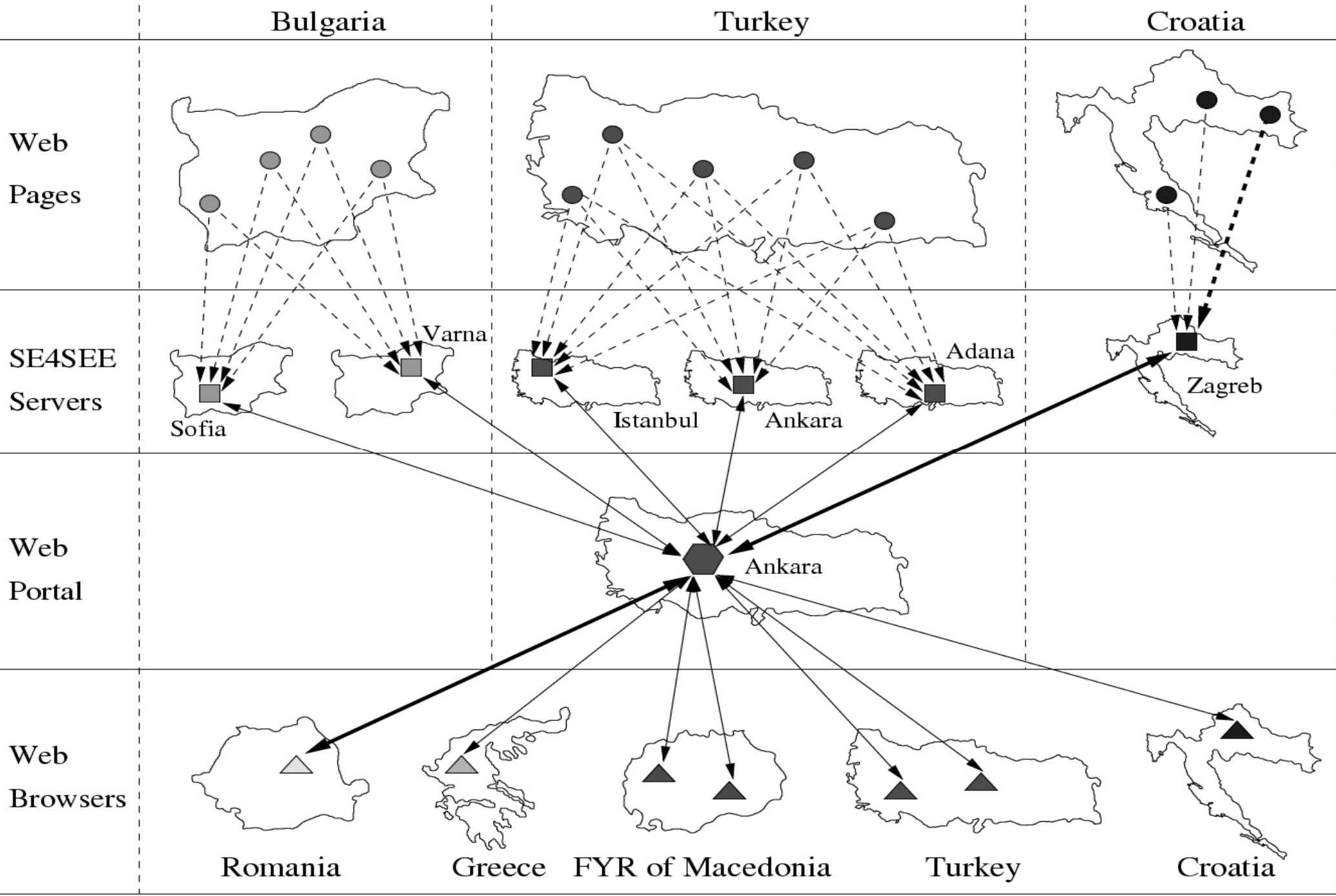
SEE-GRID
South Eastern European GRid-enabled
eInfrastructure Development

Text classification:

- Classification of Web pages under several topics
- The topics could be distinct for each country and are determined according to the characteristics of the country.
- Each country participating in the categorization phase is responsible from supplying language-dependent parts such as stop-word elimination and stemming.

The text classifier of SE4SEE is

- written in C
- is based on the Harbinger Machine Learning Toolkit
- Is currently supporting around 10 different classifiers



Differences with GRACE



SEE-GRID
South Eastern European GRid-enabled
eInfrastructure Development

GRACE

- general-purpose
- dependent on other document servers
- keyword-based search
- batch processing of queries
- categorization of query results

SE4SEE

- socio-cultural
- independent
- category-based search
- interactive
- categorization of Web pages

Benefits of a Regional Search Engine



SEE-GRID

South Eastern European GRid-enabled
eInfrastructure Development

- High page freshness due to shortened crawling cycles,
- A medium for cultural integration,
- Opportunity for running regional data mining applications over the downloaded Web content:
 - gathering information about social, cultural, political, and scientific relations between the SEE countries,
 - discovery of the folkloric elements of the countries in the region.

Conclusion



SEE-GRID
South Eastern European GRid-enabled
eInfrastructure Development

SE4SEE will hopefully

- enable country-specific search,
- provide a document collection which will allow exploiting the level of cultural dependencies,
- enhance the socio-cultural integration among the SEE countries.