

Sailing the Web with *Captain Nemo* a Personalized Metasearch Engine



(<http://www.dblab.ntua.gr/~stef/nemo>)

*Stefanos Souldatos, Theodore Dalamagas, Timos Sellis
(National Technical University of Athens, Greece)*



INTRODUCTION

Metasearching

Personalization

Metasearching & Personalization

Metasearching



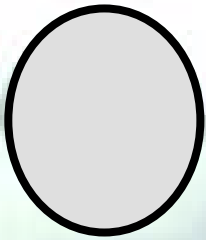
Metasearching



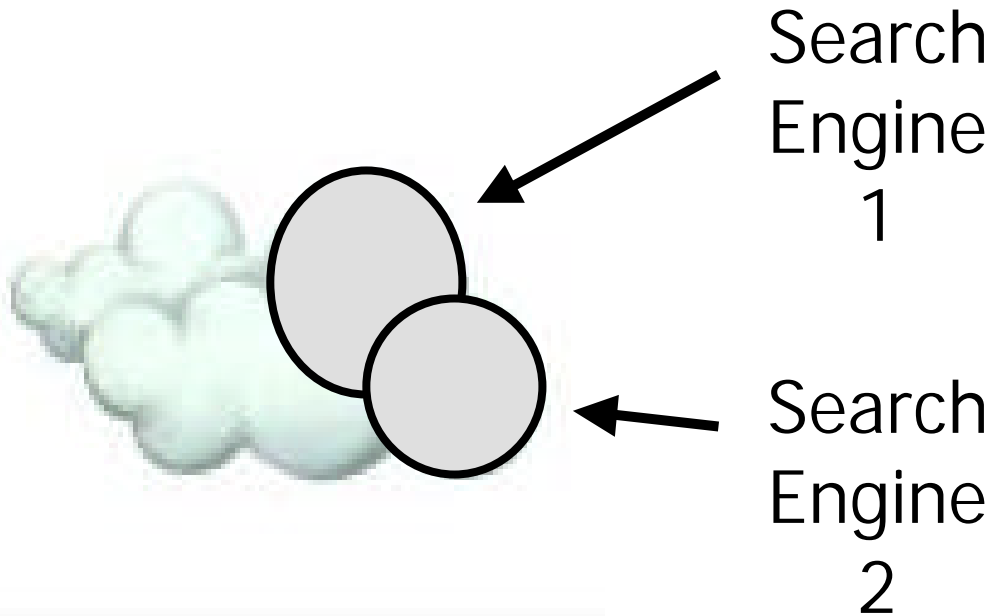
Metasearching



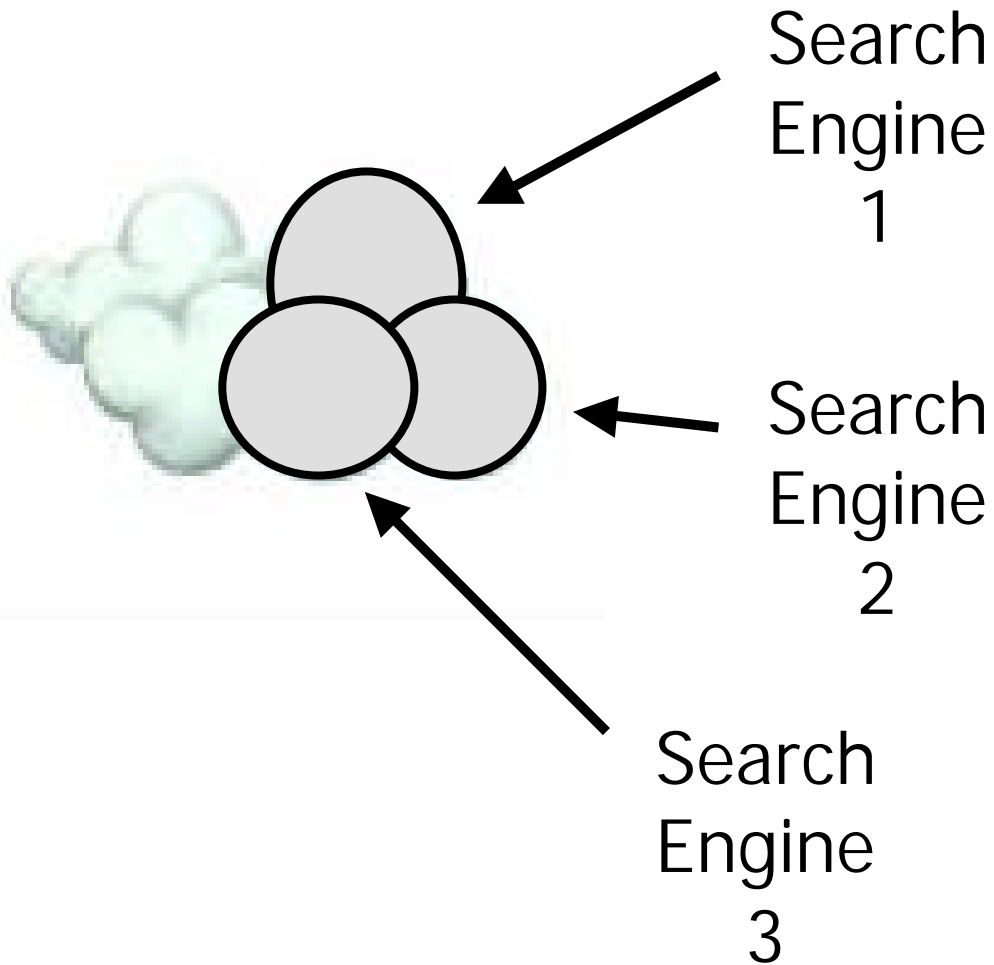
Search
Engine
1



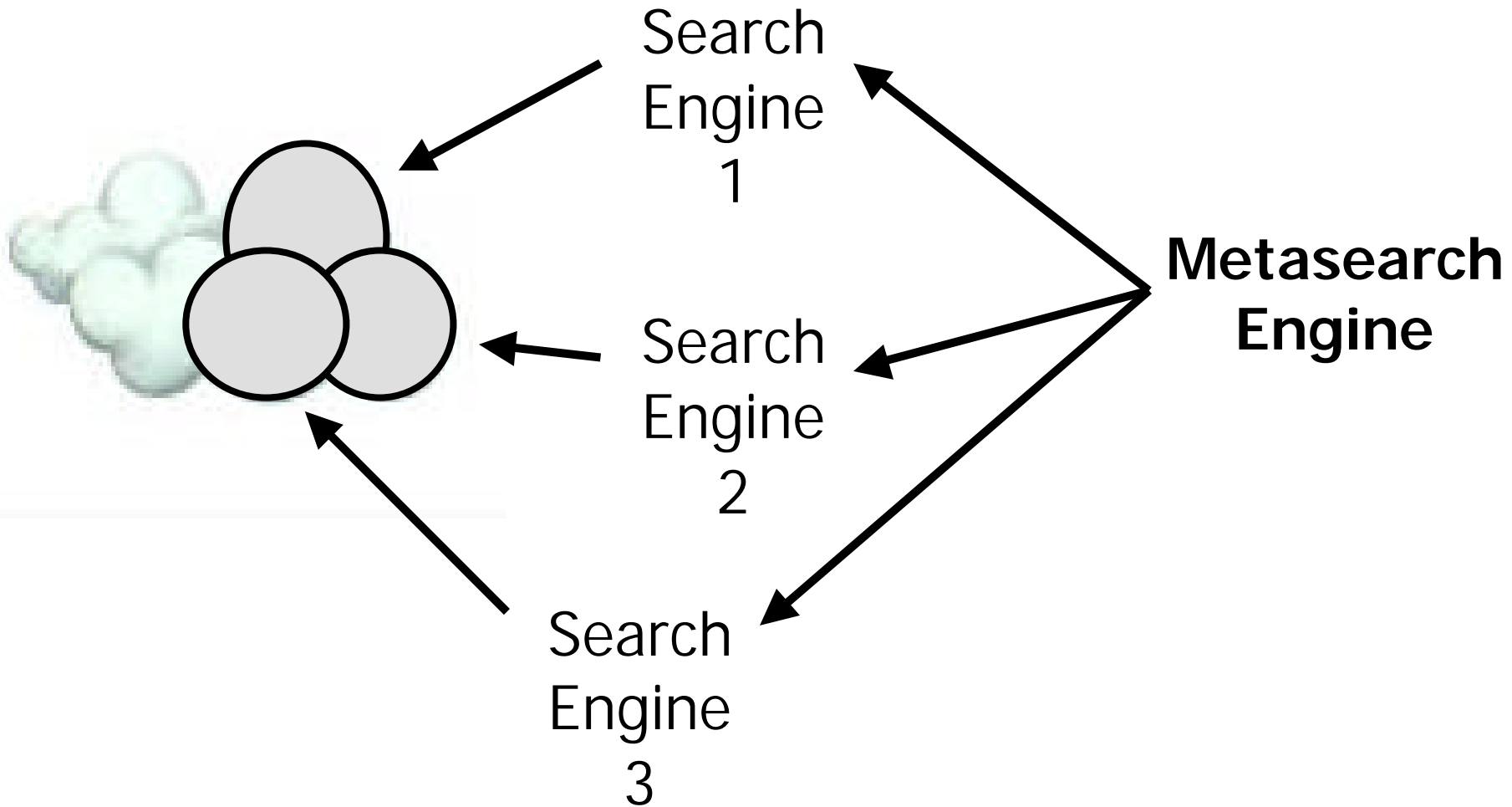
Metasearching



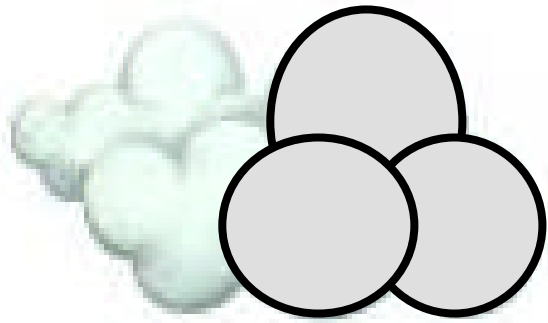
Metasearching



Metasearching



Metasearching

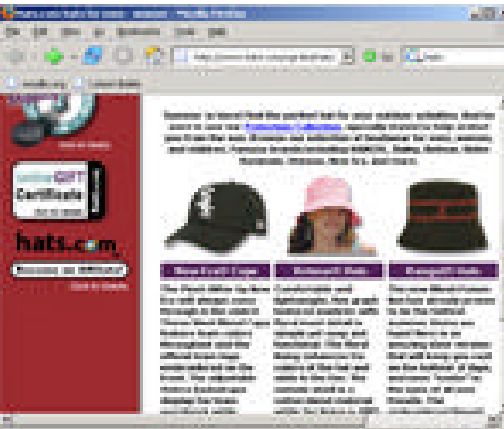


**Metasearch
Engine**

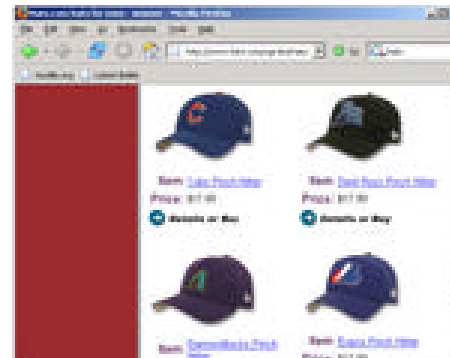
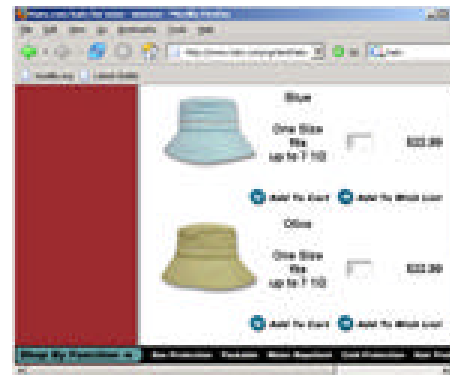
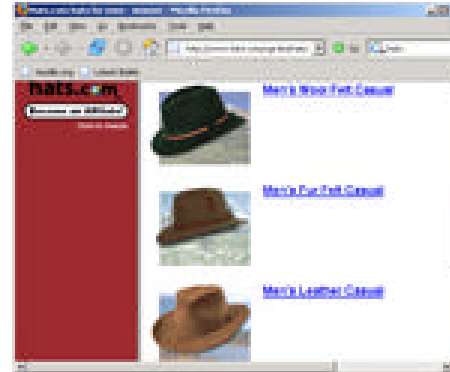
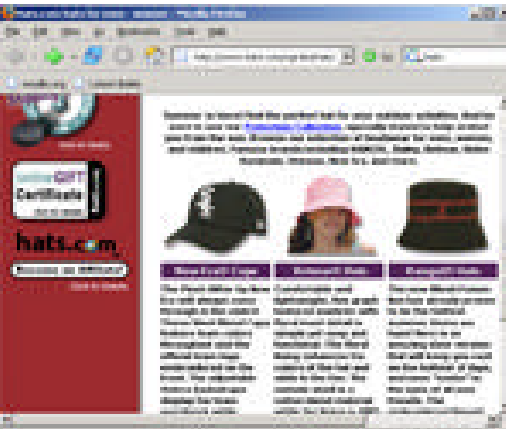
Personalization



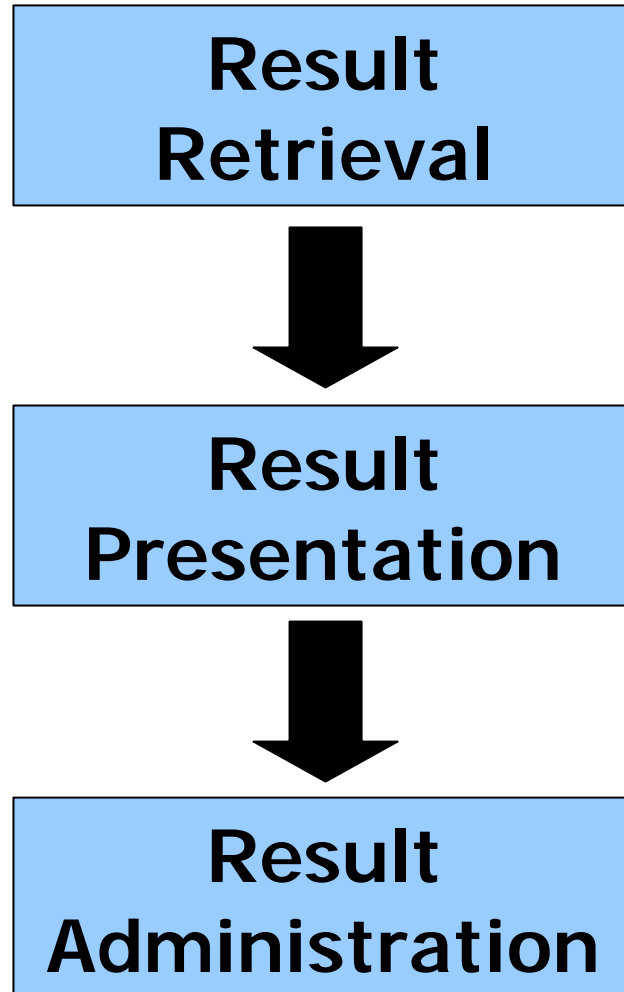
Personalization



Personalization



Metasearching & Personalization





INTRODUCTION TO *CAPTAIN NEMO*

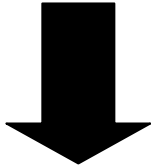
Personalization in Captain Nemo
Contribution

Personalization in Captain Nemo



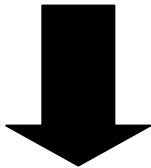
**Result
Retrieval**

**Personal Retrieval Model
(search engines, #pages, timeout)**



**Result
Presentation**

**Personal Presentation Style
(grouping, ranking, appearance)**



**Result
Administration**

**Topics of Personal Interest
(semi-automatic classification)**

Contribution



- We present personalization techniques for metasearch engines (presentation style, retrieval model, ranking algorithm).
- We suggest semi-automatic classification techniques in order to recommend relevant topics of interest to classify the retrieved Web pages.
- We present a fully-functional metasearch engine, called Captain Nemo, that implements the above framework.



RELATED WORK

Personalization in Retrieval



WebCrawler

Search

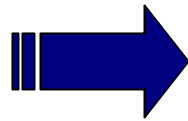
ixquick

Infogrid

Mamma

Profusion

Query Server



User defines the:

search engines to be used

timeout option (i.e. max time to wait for search engine results)

number of pages to be retrieved by each engine

Personalization in Retrieval



WebCrawler

Search

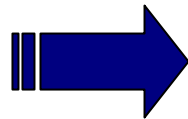
ixquick

Infogrid

Mamma

Profusion

Query Server



User defines the:

search engines to be used

timeout option (i.e. max time to wait for search engine results)

number of pages to be retrieved by each engine

Personalization in Retrieval



WebCrawler

Search

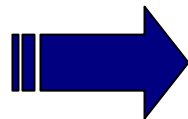
Ixquick

Infogrid

Mamma

Profusion

Query Server



User defines the:

search engines to be used

timeout option (i.e. max time to wait for search engine results)

number of pages to be retrieved by each engine

Personalization in Presentation



Search Engines

Alltheweb

Personal stylesheets to customize the look 'n' feel

AltaVista

High or low details in the description of the results

Metasearch Engines

WebCrawler

MetaCrawler

Dogpile

Result grouping by search engine that retrieved them

Topics of Personal Interest



Buntine et al. (2004)

Topic-based open source search engine

Northern Light

Organizes search results into *custom folders*

Inquirus2

Recognises categories and improves queries towards a category

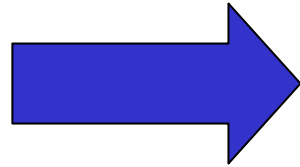
Chakrabarti et al. (1998)

Exploit link information for hypertext categorization

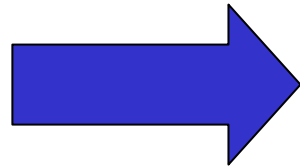


CAPTAIN NEMO

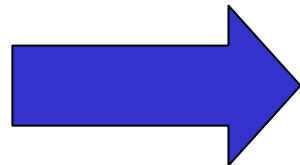
**User
Profile**



Personal Retrieval Model



Personal Presentation Style



Topics of Personal Interest

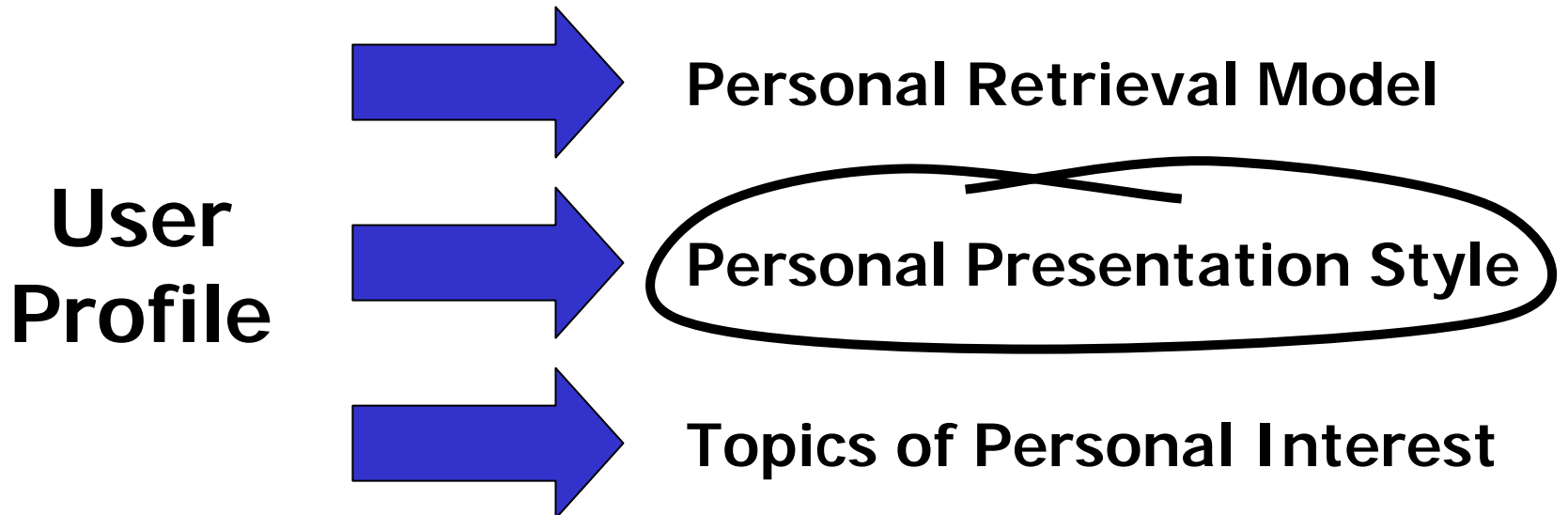
Personal Retrieval Model



	Search Engine 1	Search Engine 2	Search Engine 3
■ Search Engines	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
■ Number of Results	20	30	10
■ Search Engine Timeout	6	8	4
■ Search Engine Weight	7	10	5



CAPTAIN NEMO



Result Grouping



- Merged in a single list
- Grouped by search engine
- Grouped by relevant topic of interest

Result Content



- Title
- Title, URL
- Title, URL, Description

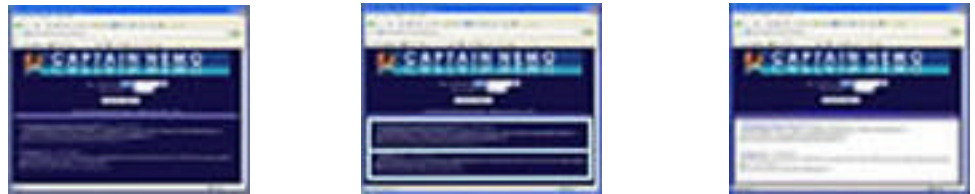
Look 'n' Feel



- Color Themes
(XSL Stylesheets)



- Page Layout

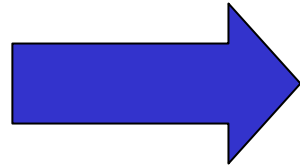


- Font Size

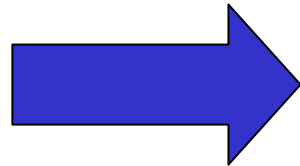


CAPTAIN NEMO

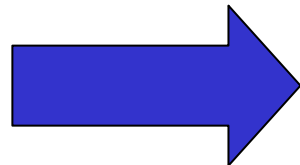
**User
Profile**



Personal Retrieval Model



Personal Presentation Style



Topics of Personal Interest

Topics Administration



- The user defines topics of personal interest (i.e. thematic categories).
- Each thematic category has a name and a description of 10-20 words.
- The system offers an environment for the administration of the thematic categories and their content.

Semi-automatic Classification



- The system proposes the most appropriate thematic category for each result.
- The user can save the results in the proposed or other category.
- The classification implements a Nearest Neighbor algorithm (Witten et al., 1999) comparing the title and description of results with the name and description of the thematic categories.

Classification Example



Topics of Interest

(t1) Sports:

football basketball
baseball swimming
tennis soccer game

(t2) Science:

scientific maths
physics computer
technology

(t3) Arts:

decorating art
painting poetry
sculpture music

Result

Alen Computer Co. can teach you the art of programming...Technology is just a game now...computer science for beginners

Classification Example



Topics of Interest

(t1) Sports:

football basketball
baseball swimming
tennis soccer game

(t2) Science:

scientific maths
physics computer
technology

(t3) Arts:

decorating art
painting poetry
sculpture music

0.287

Result

Alen Computer Co. can teach you the art of programming...Technology is just a game now...computer science for beginners

Classification Example



Topics of Interest

(t1) Sports:

football basketball
baseball swimming
tennis soccer game

0.287

(t2) Science:

scientific maths
physics computer
technology

0.892

(t3) Arts:

decorating art
painting poetry
sculpture music

Result

Alen Computer Co. can teach you the art of programming... Technology is just a game now... computer science for beginners

Classification Example



Topics of Interest

(t1) Sports:

football basketball
baseball swimming
tennis soccer game

0.287

(t2) Science:

scientific maths
physics computer
technology

0.892

(t3) Arts:

decorating art
painting poetry
sculpture music

0.368

Result

Alen Computer Co. can teach you the art of programming...Technology is just a game now...computer science for beginners

Classification Example



Topics of Interest

(t1) Sports:

football basketball
baseball swimming
tennis soccer game

0.287

(t2) Science:

scientific maths
physics computer
technology

0.892

(t3) Arts:

decorating art
painting poetry
sculpture music

0.368

Result

Alen Computer Co. can teach you the art of programming...Technology is just a game now...computer science for beginners



METASEARCH RANKING

Two Ranking Approaches



Using Initial
Scores of
Search Engines

Not Using
Initial Scores of
Search Engines

Using Initial Scores



- **Rasolofo et al. (2001)** believe that the initial scores of the search engines can be exploited.
- Normalization is required in order to achieve a common measure of comparison.
- A weight factor incorporates the reliability of each search engine. Search engines that return more Web pages should receive higher weight. This is due to the perception that the number of relevant Web pages retrieved is proportional to the total number of Web pages retrieved as relevant.

Not Using Initial Scores



- The scores of various search engines are not compatible and comparable even when normalized.
- **Towell et al. (1995)** note that the same document receives different scores in various search engines.
- **Gravano and Papakonstantinou (1998)** point out that the comparison is not feasible not even among engines using the same ranking algorithm.
- **Dumais (1994)** concludes that scores depend on the document collection used by a search engine.

Aslam and Montague (2001)



- **Bayes-fuse** uses probabilistic theory to calculate the probability of a result to be relevant to a query.
- **Borda-fuse** is based on democratic voting. It considers that each search engine gives votes in the results it returns (N votes in the first result, $N-1$ in the second, etc). The metasearch engine gathers the votes and the ranking is determined democratically by summing up the votes.

Aslam and Montague (2001)



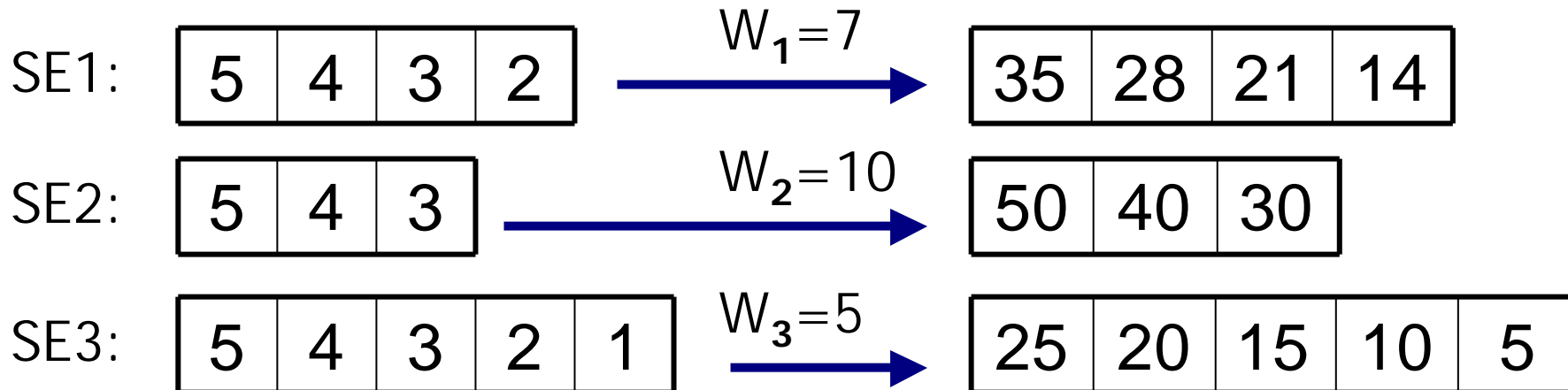
- **Weighted borda-fuse:** weighted alternative of borda-fuse, in which search engines are not treated equally, but their votes are considered with weights depending on the reliability of each search engine.

Weighted Borda-Fuse



- $V(r_{i,j}) = w_j * (\max_k(r_k) - i + 1)$
 - $V(r_{i,j})$: Votes of i result of j search engine
 - w_j : weight of j search engine (set by user)
 - $\max_k(r_k)$: maximum number of results

■ Example:





CONCLUSION – FUTURE WORK

Conclusion



- We presented Captain Nemo, a fully-functional metasearch engine with personal search spaces.
- Users can define their personal retrieval model, presentation style and topics of interest.
- Captain Nemo recommends a relevant topic of interest to classify each result, exploiting Nearest-Neighbour classification techniques.

Future Work



- To replace the flat model of topics of interest by a hierarchy of topics in the spirit of Kunz and Botsch (2002).
- To improve the classification process, exploiting background knowledge in the form of ontologies (Bloehdorn & Hotho, 2004).

Captain Nemo



<http://www.dblab.ntua.gr/~stef/nemo>

Links



- ▶ Introduction
- ▶ Introduction to Captain Nemo
- ▶ Related work
- ▶ Captain Nemo: Personal Retrieval Model
- ▶ Captain Nemo: Personal Presentation Style
- ▶ Captain Nemo: Topics of Personal Interest
- ▶ Metasearch Ranking
- ▶ Conclusion – Future Work