Search Engines

Server-Side Web Languages

Uta Priss
School of Computing
Napier University, Edinburgh, UK

Outline

Search Forms

- Minimalistic: 1 textbox, 1 button
- Detailed: "Advanced Search", many textboxes, radio buttons, menus etc

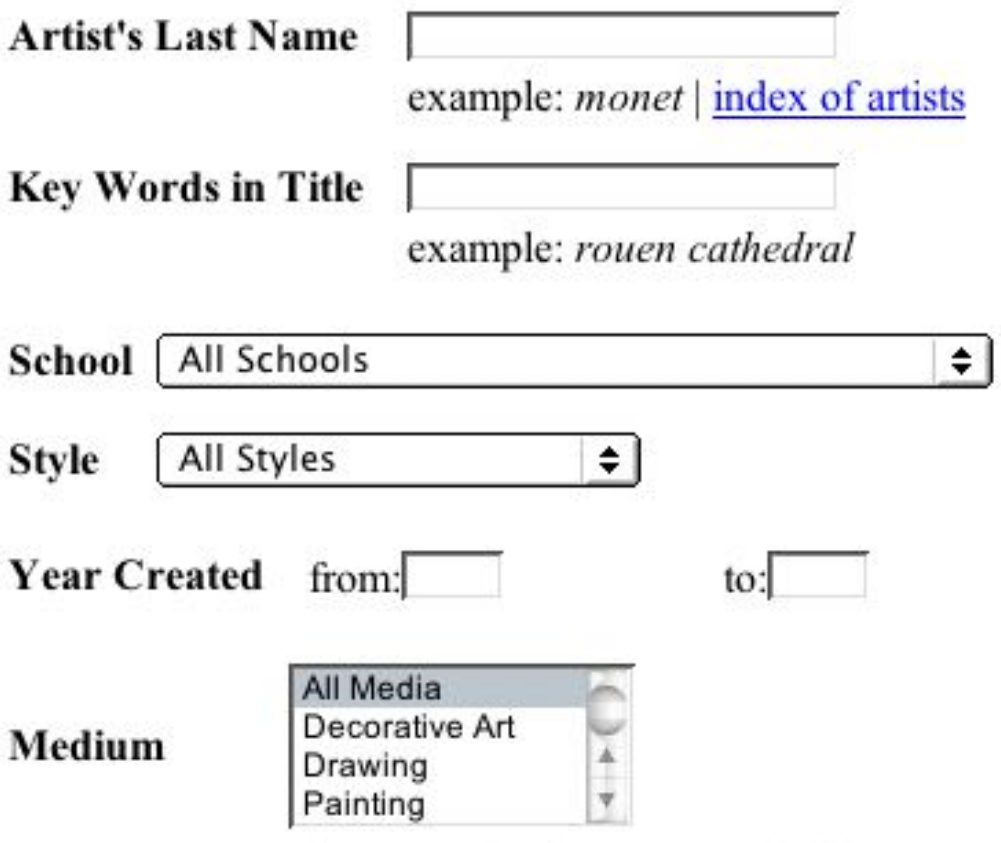
Search Forms

- Minimalistic: 1 textbox, 1 button
- Detailed: "Advanced Search", many textboxes, radio buttons, menus etc

Questions to ask: How skilled are the users?
What kind of data is queried?
How many search options will be implemented?

Example 1: Google

Example 2: National Gallery of Art



Example 3: Yahoo Advanced Search

Show results with

all of these words [ ] any part of the page

the exact phrase [ ] any part of the page

any of these words [ ] any part of the page

none of these words [ ] any part of the page

Tip: *Use these options to look for an exact phrase or to exclude pages containing search to certain parts of pages.*

Search Options

- Boolean “AND”, “OR”, “NOT”
- Search in parts of document (title, abstract, ...)
- Search using meta-data (author, date, keywords, ...)
- Combine search with categories from a directory (eg. search only documents on computer science)

Natural language equivalents to Boolean operators

AND	all words must be matched
OR	any of the words are matched
NOT	none of the words are matched

For a phrase search: "the exact phrase is matched"

The Results Screen

- Ranked list of results
- Count of results (or "No results found")
- Links to the results
- Information for each result: title, some text, URL, date, ...
- If there is a limit for how many results per page: link to previous and next sets of results
- Other features: commercial vs non-commercial results, search form, link to advanced search, help, ...

The Internals of a Search Engine

- Index: normally documents are not searched in real time. Instead an index is prepared (in form of a hash file or database), which is searched by the user.
- Stopwords: frequent words that are excluded from searches (e.g. "the", "in", "and").
- Stemming: e.g. a search for "computing" might retrieve "computer"
- Robots, spiders, crawlers: computer programs that create an index for the WWW. Can be controlled with the robots.txt file.

Algorithms for Calculating the Results

Belongs into the discipline of Information Retrieval

- Vector-space model: query space is compared to document space
- Latent semantic indexing: this is really neither "latent" nor "semantic", but a mathematical method for focusing on distinctive terms.
- Google's method: use the link structure of the web. The highest ranked results are those which are most frequently linked to.