GOOGLE FOR EDUCATIONAL RESOURCES

INFOLUNCH PRESENTATION
26/02/2003

Daniel Olmedilla
Motivation

- The number of pages in internet has been duplicated in last two years. Now there are around 1 billion. Experts think this number will be duplicated again in two years. This means that Semantic Web, crawlers and search engines will play a basic role (as now they are playing).

- A user wants to search an educational resource (e.g. a course) in a normal web search engine. He retrieves too much information and the greatest part is not related to his interests.

- In ELENA project users should join its private network to provide or query content. It requires knowledge and development. With this approach users have direct access to any public educational content in the world wide web like a normal web search engine.
Typical search engine (I)
Typical search engine (II)

Elements:
- User Interface: needed to take the user query
- Index: database/repository with the data to be searched
- Search module:
  - Transform query to understandable format
  - Do matching with the index
  - Return the results as output with information needed
- Evaluation module: take care of the user behaviour with the results of the query
Typical crawler structure (I)
Typical crawler structure (& II)

Elements:
- Retrieving module: retrieve each document from the web and give it to the Process module
- URL Listing module: has a list (ordered or unordered) of URLs and give them to the Retrieving module
- Process module: realize the next steps and after that, it gives the result (data) to the Format & Store module:
  - Automatic text analysis
  - Classification
  - Filter
- Format & Store module: convert data to better format and store it into the index
- Index: database/repository with the useful data retrieved
More information

- C.J. “Keith” van Rijsbergen, “Information Retrieval”

- Glossary for Information Retrieval
  http://www.cs.jhu.edu/~weiss/glossary.html

- Text Information Retrieval, Mining and Exploitation Fall 2002
  http://www.stanford.edu/class/cs276a/projects

- Simple crawler and search engine implementation
  http://www.learninglab.de/~olmedilla/?action=projects/buscador&topic=buscador
Why is it interesting?

- At the moment, there is not any really useful way to query this sort of documents
- A specific method to rank results of educational resources does not exists yet
- Semantic Web approaches:
  - Insert embedded metadata into html page
    - Problem: it is very difficult to change 1 billion web pages and there are many inexpert users
  - Semi-automatic annotation for the entire WWW
    - Problem: difficult to classify the whole content of WWW. There is no standard yet
- This approach:
  - No need for changes into web pages
  - Focused in only one field so only filtering is needed (not classification)
  - Solution between current search engines and Semantic Web approaches
- If this work success, it could be extent to any other field: medicine, science... (multipurpose search engine?, focused classification techniques?)
First idea to solve it

Specific crawl for learning and educational resources:
- How to distinguish between this sort of resources and the rest?
  - Similarity function + text analysis → how similar is one web page against some examples we had?
  - Neural network + adequate training → find the right parameters and train the network with known educational web pages
  - Links structure analysis → if you have an university page with words like “course”, probably each link in this page will be an educational resource or a pointer to a page which contains educational resources
- Start into specific part of the world wide web (e.g. German and Spanish universities web pages)

Specific ranking algorithms and classification
- Google rank the results in order to how much is an URL referenced in the rest of web pages (PageRank importance metric)
- Start with this approach and think and test new ones
Next steps

- Look for an already implemented crawler → In Process
  - WebBase → Done
  - Kaon → In process

- Download locally for testing pages of universities in Germany and Spain. Use WebBase for this purpose → Done
  - 10 levels of deep or 20,000 web pages crawling each site
  - 113 universities crawled
  - ≈ 850,000 pages crawled
  - ≈ 2.5 GB compressed text crawled

- Modify the Process module to filter only educational resources
- Test filtering
- Add ranking algorithms for the output results of searching
- Test ranking
- Add feedback evaluation to the ranking algorithms would be interesting to study
Workplan (I)

WebBase architecture and features → Done
- Parallel crawlers
- Page selection
- Page refresh policy
- Crawler architecture

WebBase crawler
- German universities web sites crawled → Done
- Spanish universities web sites crawled → Done

WebBase client → In process
- Shell queries enabled
  - Stream queries → Done
  - Queries by URL
- Java API query → Not developed yet
Workplan (II)

Introduction to source code
- C++ classes and roles
  - Crawler → In process
  - Client

WebBase final report → In process
- Installation (database, crawler and client) → Done
- Execution (database, crawler and client) → Done
- Indexing → In process
- Results

Infolunch presentation at L3S
- Brief resume of work done → In process ;-)

Daniel Olmedilla 3/27/04
Useful links

- Stanford WebBase project
- Citeseer
  [http://www.neci.nec.com/~lawrence/pub-ri.html](http://www.neci.nec.com/~lawrence/pub-ri.html)
- Stanford Database Group Publication Server
  - Searching the web
  - Topic-Sensitive PageRank
  - The anatomy of a Large-Scale Hypertextual Web Search Engine
- KAON: The Karlsruhe Ontology and Semantic Web Tool Suite