Subject-based information organization: KnowLib's findings

Koraljka Golub, Knowledge Discovery and Digital Library Research Group
http://www.it.lth.se/knowlib/

LIVA meeting at BTJ, 5 April 2006

Outline

- Subject browsing
- Automated subject classification
- Focused crawling
- Demonstrators

Subject browsing

- seeking for information resources by examining a hierarchical tree of broader and narrower subject classes into which the resources have been classified
- browsing services
  - for academic users
    - e.g. Renardus (http://www.renardus.org)
  - commercial
    - e.g. Google Directory (http://www.google.com/dirhp)
- browsing vs. searching
  - contradictory claims and research results

Structures for subject browsing

- traditional: classification schemes, thesauri, subject heading systems
- from the WWW: ontologies, search-engine directories
- some better for browsing than others
  - hierarchical structure
  - document collection
  - names of subjects

Renardus

- http://www.renardus.org
- integrated searching and browsing of ca. 80000 resources from major European subject gateways
  - simple and advanced searching
  - browsing through Dewey Decimal Classification (DDC)
  - browsing support features

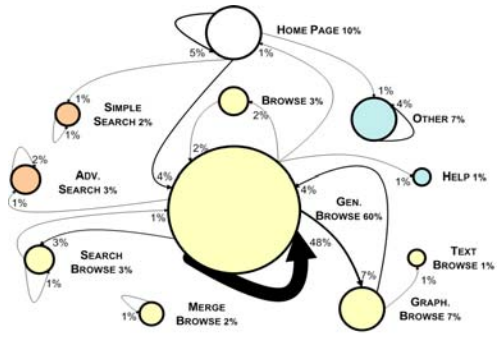
Research issues

- the balance between browsing, searching and mixed activities
- the degree of usage of the browsing support features
- typical sequences of user activities and transition probabilities in a session, esp. in traversing the hierarchical DDC browsing structure
- typical entry points and referring sites

Methodology

- log analysis
  - users do not need to be directly involved
  - catches unsupervised behaviour
  - every activity within the system tracked
  - cleaned and categorized entries (ca. 460000) grouped into user sessions (ca. 73000)
    - all entries from the same address
    - time gap between two entries less than 1 hour
    - one-entry sessions & sessions shorter than 2 seconds removed
- sample
  - 16 months (2002/2003)

Main activities and transitions



Dominance of browsing

- 76% of all activities are browsing
  - majority start using Renardus at a browsing page because directly referred by a search engine
  - layout of Home page "invites" browsing
    - also users starting at Home page predominantly use browsing
- good usage of browsing support features, esp.:
  - graphical overview
  - search entry to browsing pages
- 5% of all activities are searching

Two types of users

- 71% people referred by search engines (mostly Google and Yahoo!)
  - 87% browsing, 2,7% searching
- 22% start at Home page
  - 57% browsing, 12,5% searching
  - more browsing activities per session than the other type
  - use non-browsing activities 3x (Other) and 5x (searching) as often
  - have 2x as many activities per session (ca. 10)
  - they use the service elaborately, in a way system designers intended

DDC browsing

- 60% of all activities
- 2/3 are in unbroken browsing sequences
  - up to 86 steps
- keywords
  - good chance of finding browsing pages when using more than one search terms

Major results

- given proper conditions, browsing is heavily used
  - browsing support features are also heavily used
- it is implied that DDC could serve as a good browsing structure, including terminology

Outline

√ Subject browsing

→ Automated subject classification

❖ Focused crawling

❖ Demonstrators

Automated subject classification

- subject classification
  - grouping documents that have a property (topic, theme) in common, further sub-grouping of documents based on finer properties
  - establishing relationships between them
- automated subject classification
  - machine-based (statistical, NLP techniques)
  - approaches
    - text categorization
    - document clustering
    - document classification

Text categorization

- machine learning
  - algorithms
- information retrieval
  - vector-space model
  - evaluation measures
- pre-defined browsing structures
  - learning about categories from pre-existing documents in the categories
  - for Web pages, search-engine directories

Document clustering

- information retrieval
- vector-space model
- browsing structures automatically derived
  - clusters of similar documents and, partially, relationships between them
  - names of the clusters
  - such structures hard to understand
  - rather unstable as well

Document classification

- library science approach
- pre-defined browsing structures
  - controlled vocabularies, usu. classification schemes
  - good for browsing
- no vector representations
  - string-to-string matching against a controlled vocabulary

Mixed approach

- text categorization or information retrieval algorithms
- controlled vocabularies with structures well suited for browsing (usu. classification schemes, not search-engine directories)
- few examples

Issues

- automating subject determination
  - logical positivism
    - subject is a string occurring a certain number of times, in a certain location etc.
    - if document 1 is about subject A, and if document 2 is similar to document 1, then document 2 is also about subject A
- evaluation
  - issue of deriving the correct interpretation of a document's subject matter
  - few end-user evaluations

Similarities between approaches

- document pre-processing and indexing
  - removing stop-words
  - extracting relevant words
- utilization of text-document characteristics
  - structural elements
  - metadata
  - text neighbouring headings and anchor text
  - text from linked pages
- assumption: idea exchange beneficial

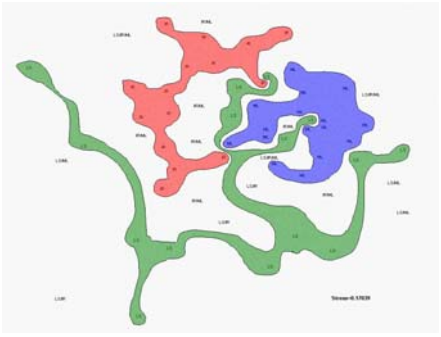
Is there an exchange of ideas?

- main research question
  - to what degree the three communities utilize others' ideas, methods, and findings
- direct links
  - do authors from one community cite authors from another
- indirect links
  - bibliographic coupling of papers
- sample
  - 148 papers: 52 ML, 63 IR, 33 LS

Direct links

- the ML community uses IR methods and both tended to cite each other to a certain extent
- few cases where LS authors were cited by either of the two other communities and the other way around

Indirect links



Major results

- on the sample of 148 papers, it was shown that the three communities dealing with automated classification of Web pages do not communicate to a large extent
- there is a more evident link between machine learning and information retrieval communities
- library science community is rather isolated

Comparing approaches

Class	String-matching			Machine learning (SVM)		
	Rec.	Prec.	F1	Rec.	Prec.	F1
402	0.58	0.49	0.53	**0.93**	**0.91**	0.92
722.3	0.12	0.26	0.16	**0.76**	**0.79**	0.78
723.1.1	0.34	0.32	0.33	**0.74**	**0.79**	0.76
723.4	0.37	0.39	0.38	**0.65**	**0.81**	0.72
903	0.28	0.61	0.38	**0.72**	**0.74**	0.73
903.3	0.32	**0.97**	0.48	**0.74**	0.79	0.76
Microavg.	0.55	0.68	0.61	**0.78**	**0.81**	**0.78**
Macroavg.	0.55	0.68	0.61	**0.76**	**0.81**	**0.78**

Using Web-page elements

- what is the importance of distinguishing between different parts of a Web page?
  - title, headings, main text, metadata
  - what are the appropriate significance indicators?

e.g. http://froggy.lbl.gov/virtual/

<title>Virtual Frog Dissection Kit Version 2.2</title>
<meta name="description" content="Virtual Frog Dissection Kit">
<meta name="keywords" content="frog dissection K-12 education">
<h2 align="center">Virtual Frog Dissection Kit</h2>
<h2>Frog watch</h2>
main text:
"This award-winning interactive program is part of the "Whole Frog" project. You can interactively dissect a (digitized) frog named Fluffy, and play the Virtual Frog Builder Game. The interactive Web pages are available in a number of languages...."

Structural elements and metadata

- collection
  - 1003 Web pages in engineering
  - Ei classification scheme
    - 6 main classes
    - decimally subdivided
    - up to 5 hierarchical levels

4 Civil Engineering
...
44 Water and Waterworks Engineering
441 Dams and Reservoirs
...
445 Water Treatment
445.1 Water Treatment Techniques
445.1.1 Potable Water Treatment Techniques
...

Approach

- algorithm
  - when a match is found, the corresponding class is assigned, with a relevance score, based on:
    - which term is matched (single word, phrase, Boolean)
    - type of class matched (main or optional)
    - the part of the Web page in which the match is found
- significance indicators
  - derived using various measures of correctness
    - precision and recall
    - semantic distance
    - multiple regression

Major results

- title performs best, followed by headings, metadata, and text
- necessary to use all structural elements and metadata (not all of them occur on every Web page)
- how to combine them not important
  - the best combination was only 3% better than the worst one

Improving classification

- termlist expansion
  - syntactic expansion
  - semantic expansion
    - manual, machine learning, NP extraction
- adjusting term weighting
- adjusting cut-off

Outline

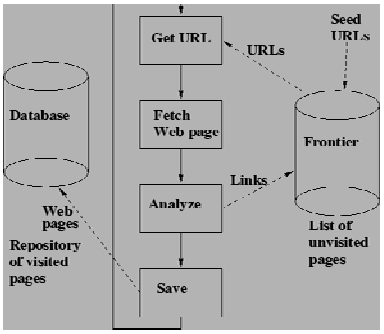
√ Subject browsing

√ Automated subject classification

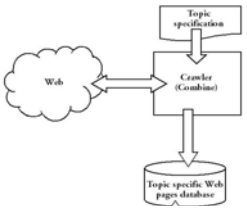
→ Focused crawling

❖ Demonstrators

Simple crawling



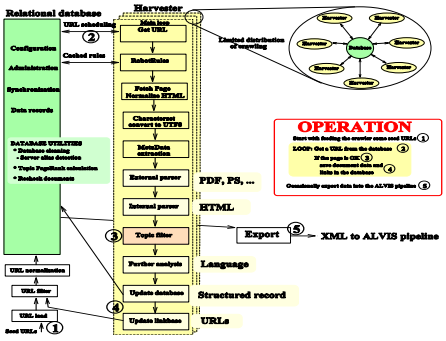
Focused crawling in ALVIS



focused crawling in ALVIS:
http://www.it.lth.se/knowlib/publ/ESWC.xfig.v4.pdf

ALVIS: http://www.alvis.info

Focused crawling in ALVIS



Combine focused crawler

- availability: http://combine.it.lth.se/
  - download, documentation, publications)
- testbed databases
  - Materials science (1 650 000 records)
  - Bacillus subtilis (55 000 records)
  - Search engines (700 000 records)
  - Carnivorous plants (80 000 records)
  - Engineering (600 000 records)
  - Malaria (85 000 records)

Outline

√ Subject browsing

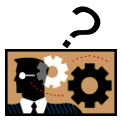
√ Automated subject classification

√ Focused crawling

→ Demonstrators

Demonstrators

- http://www.it.lth.se/knowlib/demos.htm
  - also, automatic vocabulary mapping http://dbkit02.it.lth.se/exp/map/



&