Automated subject classification of textual Web pages, for browsing

Koraljka Golub
KnowLib, Department of Information Technology, Lund University
http://www.it.lth.se/knowlib/

Practical info

- doctoral project period: July 2003 – July 2007
- thesis form: compilation of papers
  - 2 different kinds of dissertations in Sweden
    1. monographs
       - a unified and coherent work
       - almost exclusively in the humanities, theology and law
    2. compilations
       - 3-6 peer-reviewed papers published during the period of postgraduate training and a summary of the papers
       - 2/3s of all theses in Sweden

(From http://www.doktorandhandboken.nu/english/phd_studies.html#3_3)

Purpose

- explore to what degree automated subject classification based on a controlled vocabulary could be utilized in automated subject classification of textual Web pages
  - in comparison to state-of-the-art approach to text categorization, SVM, which requires training documents
  - the context of browsing
  - user-based evaluation

Research questions

- Is hierarchical browsing of Web pages being used today?
- How does performance of SVM compare to performance of the string-matching algorithm on a test collection?
- What can be done to improve performance of the string-matching algorithm on a test collection?
- How does performance of SVM compare to performance of the string-matching algorithm, in the context of browsing of Web pages?

Approaches to classification

- SVM
  - machine-learning
  - state-of-the-art algorithm for automated text classification
  - requires training documents
    - cca. 8000 training and testing documents per class in RCV1
- String-matching on controlled vocabulary
  - matching of strings in text to be classified against terms designating subject classes in a controlled vocabulary
  - so far in research considered too simple to be good
    - our assumption: a good controlled vocabulary provides enough for good classification performance
  - doesn't require training documents

Lacks in current research

- evaluation challenge
  - automated classification performance not really tested in the context of operational systems and users
    - main problem: indexing inconsistency
- automated classification has not been tested in the context of browsing

Controlled vocabulary

- 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

  - pre-existing intellectual mappings between terms in the Ei thesaurus to terms in the Ei classification scheme

Table 1. The number of different types of terms for 92 sub-classes from class 9

	All	BT	Ca	NT	PT	RT	ST
Total	8099	932	92	1423	1691	4378	1739
Avg./class	88	10	1	15	18	48	19

Problem: test collection

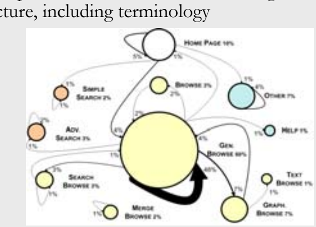
- test collection of Web pages
  - cca. 1000 Web pages
    - the only one that there is for the chosen controlled vocabulary
    - very small compared to standard test collections for automated classification (Reuters RCV1 has cca. 100 classes and 800,000 documents)
  - decision:
    1) conduct performance evaluation of the SVM algorithm on scientific abstracts collection against intellectually assigned classes
    2) conduct performance evaluation of string-matching algorithm based on different parameters on the same scientific abstracts collection against intellectually assigned classes
    3) crawl a collection of Web pages, classify them with each of the two algorithms and conduct user-based evaluation

Methodology so far

- log analysis of Renardus for user behaviour
- 92 classes selected from the class Engineering, General (class 9)
  - about 35,000 abstracts from Compendex
- the 1000 Web page collection
  - used for deriving weights for different parts of Web pages

Browsing of Web pages

- browsing of Web pages is heavily used when using classification schemes such as DDC, to which Ei is similar
- it is implied that DDC could serve as a good browsing structure, including terminology



Koch, T., Golub, K., and Ardö, A. 2005. Users browsing behaviour in a DDC-based Web service: A Log Analysis. Accepted for publication in Cataloging & Classification Quarterly.

SVM vs. string-matching

- on the abstracts collection, SVM outperforms string-matching
  - string-matching in its simplest form without weighting and cut-offs

Table 2. Experimental results comparing performance of the two approaches per class, and number of original terms per class in SM (Terms).

	String matching (SM)				Machine learning (SVM)		
Class	Terms	Recall	Precision	F1	Recall	Precision	F1
402	423	0.58	0.49	0.53	**0.93**	**0.91**	0.92
722.3	292	0.12	0.26	0.16	**0.76**	**0.79**	0.78
723.1.1	137	0.34	0.32	0.33	**0.74**	**0.79**	0.76
723.4	61	0.37	0.39	0.38	**0.65**	**0.81**	0.72
903	58	0.28	0.61	0.38	**0.72**	**0.74**	0.73
903.3	26	0.32	**0.97**	0.48	**0.74**	0.79	0.76
Microavg.		0.35	0.45	0.39	**0.78**	**0.81**	**0.78**
Macroavg.		0.34	0.51	0.38	**0.76**	**0.81**	**0.78**

Golub, K., Ardö, A., Mladenic, D., Grobelnik, M. 2006. Comparing and Combining Two Approaches to Automated Subject Classification of Text. In J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, Spain. P. 467-470.

Improving string-matching

- words from all elements of Web pages need to be taken, but it doesn't really matter which weight you use
- the best results in F1 were 3% better than baseline:

Score = 86*ScoreTitle + 5*ScoreHeadings + 6*ScoreMetadata + ScoreMainText .

Golub, K., and Ardö, A. 2005. Importance of HTML structural elements and metadata in automated subject classification. In: Research and Advanced Technology for Digital Libraries, Proceedings of ECDL 2005 – the 9th European Conference on Research and Advanced Technology for Digital Libraries, Vienna, Austria, 18-23 September 2005. P. 368-378.

Currently working on

- improving string-macthing
  - weights and cut-offs
  - NLP applied to string-matching
    a. single-word inflection
    b. single-word derivation
    c. multi-word morpho-syntactic analysis (c.1. change order, c.2. derivation and permutation, c.3. coordination, c.4. insertion)
    d. semantic variation (d.1. synonymy, d.2. hyperonymy, d.3. manual verification)

To do: user study

- create test collection
  - harvest Web pages from engineering
    - e.g. from Scirus or EEVL
- classify them using the two algorithms
  - SVM's training documents are Compendex, since no Web pages classified into Ei
- purpose of the user study:
  1. classification accuracy – topical relevance
  2. browsing – being able to find the right class

Issues to discuss

- Any problems with research done so far?
- How to design the final user study?
  - suggestions
    - create a simple user interface
    - given a task such as "find documents on irrigation":
      - first, find the right class
      - second, in that class look at the documents and say if they are in the right place
  - how to make sure to evaluate only accuracy and browsing, and not user interface etc.
  - how many and what type of tasks for how many participants in the study