Controlled-vocabulary based approach to automated subject classification of textual Web pages

Abstract

The purpose of this project is to explore the role of controlled vocabularies such as thesauri and classification schemes in automated subject classification of text. Apart from for improving classification performance, controlled vocabularies have been used in information retrieval systems to improve information retrieval, which is the application context of this project. The classification algorithm comprises string-to-string matching between words in the documents to be classified and words in term lists derived from the controlled vocabulary. The advantage of using this type of algorithm is that no training documents are required and, unlike in document clustering, an appropriate, good-quality controlled vocabulary can be chosen.

The chosen test controlled vocabulary is Engineering Information (Ei) classification scheme, which has mappings to the corresponding Ei thesaurus. Intended end-users are engineering students and other subject experts. Evaluation would be performed at two main levels: comparison of automatically- against manually-assigned classes, and information retrieval relevance assessments.

The project includes the following research questions: to what degree different types of terms in Ei thesaurus and classification scheme influence automated classification performance, enriched with their inflected, derivated, permuted forms as well as synonyms and hyperonyms, and comparison of this algorithm with an SVM and a clustering algorithm.

1 Introduction

Automated classification has been a challenging research issue for several decades now. A major motivation has been high costs of manual subject classification, in terms of time and human resources. Due to the ever-increasing number of documents, there is a danger that recognized objectives of bibliographic systems (finding, collocating, choice, acquisition, navigation) (Svenonius 2000, p. 20-21) would get left behind; automated means could be a solution to preserve them (ibid. p. 30).

According to K. Golub (2006), one can distinguish between three major approaches to automated classification, the biggest being text categorization, document clustering and document classification, While the first two approaches use complex algorithms, they by tradition hardly utilize controlled vocabularies. The latter focuses less on algorithms and more on operational systems using controlled vocabularies. That approach is more or less based on string-to-string matching of controlled vocabulary terms and text in documents to be classified. Usually weighting schemes are applied with the purpose of indicating degrees to which a term from a document to be classified is significant for the document's topicality. The major advantage of this approach in comparison to the other two is that no training documents are required. Controlled vocabularies (such as classification schemes, thesauri, subject heading systems) have the devices to control polysemy, synonymy, and homonymy of the natural language, and as such could serve as good-quality structures for subject searching and browsing. Another motivation to apply this approach is to re-use the intellectual effort that has gone into creating such a controlled vocabulary. For further details on the advantages of using pre-existing controlled vocabularies as well as on different approaches to automated classification and indexing see K. Golub (2006), M.-F. Moens (2000).

String-to-string matching has been explored in linguistics, and controlled vocabularies have been used in automated subject indexing. However, controlled vocabularies largely differ from one another as to their suitability for the task of automated classification or indexing, especially since they have been traditionally designed for other tasks. To the author's knowledge, Engineering Information thesaurus and classification scheme (Ei thesaurus 1995) has not been explored in this specific respect by others. In addition, the documents that have been mostly dealt with in these two areas were more traditional document forms, such as research papers, news articles etc., and not Web pages. Web pages have specific characteristics such as hyperlinks and anchors, metadata, and structural information, all of which could serve as complementary features to improve automated classification. On the other hand, they are rather heterogeneous; many of them contain little text, metadata provided are sparse and can be misleading, structural tags can be misused, and titles can be without any information significant of the content (e.g. 'Home Page', 'Untitled Document').

2 Algorithm

The basic algorithm is based on an automated classification approach (Koch and Ardö 2000) that has been developed within the DESIRE project (DESIRE 2000).

The algorithm classifies textual documents into classes of the Ei classification system. Mappings exist between the Ei classes and Ei thesaurus' descriptors; both the captions of classes and the descriptors are matched against extracted title, headings, metadata, and main text of a Web page. A list of suggested classes and corresponding relevance scores (S) is produced using the following algorithm:

$$S = \sum_{\text{locs}} \left(\sum_{\text{terms}} \left(freq[loc_j][term_i] * weight[term_i] * weight[loc_j] \right) \right) .$$

Term weight (*weight[term$_i$])* is taken from the term list (see 3.2). Location weight (*weight[loc$_j$]*) is defined for locations like title, metadata, HTML headings, and main text. The applied formula was 86*scoreTitle, 5*scoreHeadings, 6*scoreMetadata, 1*scoreText, as determined in K. Golub and A. Ardö 2005. Frequency (*freq[loc$_j$][term$_i$]*) is the number of times *term$_i$* occurs in the text of location *loc$_{j.}$*

Only classes with scores above a pre-defined cut-off value are selected as the classes for the document.

3 Proposed research

3.1 Research questions

3.1.1 To what degree could the following elements of Ei improve automated subject classification of textual Web pages: captions, hierarchical structure, thesaurus terms and relationships between terms (e.g., related, narrower or broader). This would also include enriching Ei with the terms' synonyms and hypernyms, and different forms (e.g. single-word inflection, single-word derivation, multi-word morpho-syntactic analysis such as change of order, derivation and permutation, coordination, insertion).

3.1.2 How does the best classification performance gained using the string-to-string matching compare to an SVM (Support Vector Machine) algorithm, and a clustering algorithm.

3.2 Methodology

Test collection. The test collection to be used for developing the classification algorithm should have a sufficient number of textual documents and metadata describing their content. Each metadata record should contain manually assigned subject class from Ei. Since there do not seem to be any Web-page collections classified using Ei, the algorithmic evaluation would be conducted on research article collection Compendex. Then a selection of Web pages would be classified using the approach that performed best when tested on Compendex, and a sample would be evaluated by subject experts (see subsection on Evaluation).

Variations. A number of parameters will need to be investigated, such as:

1. Which words to include in a stop-word list;
2. Which weights to assign to extracted terms, e.g., based on tf*idf measure;
3. Which cut-off values to apply.

Evaluation. Precision, recall and F1 measure will be used as standard evaluation measures (Moens 2001, p.104-105).

Three evaluation methods will be used:

1. Comparison of automatically assigned classes against the manually assigned ones (only possible for the article collection from Compendex). Different levels of matching could be tested, e.g.:
    - total match, e.g., if the class "932.2.1." is the correct one, than the one automatically assigned needs to look exactly the same;
    - partial match, the first three digits, e.g., "932.2.1." and "932.2." have the same first three digits;
    - partial match, the first two digits, e.g., "932" and "933" have the same first two digits.
2. A sample of both manually-assigned and automatically-assigned classes will be evaluated also by subject experts.
3. Task-based task-based evaluation and/or evaluation using relevance assessments will be also conducted.

References

1) DESIRE: Development of a European Service for Information on Research and Education. Available online at: http://www.desire.org/ (accessed 22 December 2004)
2) Golub, K. Automated subject classification of textual Web documents. Journal of Documentation, 62,3(2006), 350-371.
3) Golub, K., and Ardö, A. Importance of HTML structural elements and metadata in automated subject classification, in Proceedings of ECDL 2005 – the 9th European conference on research and advanced technology for digital libraries, 368-378.
4) Koch, T., and Ardö, A. Automatic classification of full-text HTML-documents from one specific subject area. EU Project DESIRE II D3.6a, Working Paper 2. 2000. Available online at: http://www.lub.lu.se/desire/DESIRE36a-WP2.html (accessed 13 January 2006).
5) Moens, M.-F. Automatic indexing and abstracting of document texts. Boston: Kluwer, 2000.
6) Svenonius, E. The intellectual foundations of information organization. MIT Press, Cambridge, MA, 2000.