Importance of HTML Structural Elements and Metadata in Automated Subject Classification

Koraljka Golub and Anders Ardö

Knowledge Discovery and Digital Library Research Group (KnowLib),
Digital Information Systems, Department of Information Technology, Lund University ,
P.O. Box 118, 22 100 Lund, Sweden
{anders.ardo, koraljka.golub}@it.lth.se
http://www.it.lth.se/knowlib/

Abstract. The aim of the study was to determine how significance indicators assigned to different Web page elements (internal metadata, title, headings, and main text) influence automated classification. The data collection that was used comprised 1000 Web pages in engineering, to which Engineering Information classes had been manually assigned. The significance indicators were derived using several different methods: (total and partial) precision and recall, semantic distance and multiple regression. It was shown that for best results *all* the elements have to be included in the classification process. The exact way of combining the significance indicators turned out not to be overly important: using the F1 measure, the best combination of significance indicators yielded no more than 3% higher performance results than the baseline.

1 Introduction

Automated subject classification has been a challenging research issue for several decades now, a major motivation being high costs of manual classification. The interest rapidly grew around 1997, when search engines couldn't do with just full-text retrieval techniques, because the number of available documents grew exponentially. Due to the ever-increasing number of documents, there is also a danger that recognized objectives of bibliographic systems (finding, collocating, choice, acquisition, navigation) ([19], p. 20-21) would get left behind; automated means could be a solution to preserve them (ibid., p. 30). Automated subject classification of text finds its use in a wide variety of applications, such as: organizing documents into subject categories for topical browsing, which includes grouping search results by subject; topical harvesting; personalized routing of news articles; filtering of unwanted content for Internet browsers; and, many others (see [17], [12]).

A frequent approach to Web-page classification has been a bag-of-words representation of a document, in which all parts of a Web page are considered to be of equal significance. However, unlike other text documents, Web pages have certain characteristics, such as internal metadata, structural information, hyperlinks and anchors, which could serve as potential indicators of subject content. For example, words from title could be more indicative of a page's content than headings. The degree to which

A. Rauber et al. (Eds.): ECDL 2005, LNCS 3652, pp. 368 – 378, 2005.

different Web page elements are indicative of its content is in this paper referred to as significance indicator.

With the overall purpose of improving our classification algorithm (see section 2.3), the aim was to determine the importance of distinguishing between different parts of a Web page. Significance of four elements was studied: title, headings, metadata, and main text.

The paper is structured as follows: in the second chapter a literature review is given, evaluation issues are discussed and the algorithm used is described (2 Background); in the third chapter data collection as well as methodology for deriving significance indicators are described (3 Methodology); deriving and testing the significance indicators is presented in chapter 4 (4 Significance indicators). The paper ends with conclusions and further research (5 Conclusion).

2 Background

2.1 Related Work

A number of issues related to automated classification of documents and significance of their different parts have been explored in the literature. A. Kolcz, V. Prabakarmurthi, J. Kalita and J. Alspector [14] studied news stories features and found out that initial parts of a story (headline and first two paragraphs) give best results, reflecting the fact that news stories are written so as to capture readers' attention. J. Pierre [16] gained best results in targeted spidering when using contents of keywords and description metatags as the source of text features, while body text decreased classification accuracy. R. Ghani, S. Slattery & Y. Yang [10] also showed that metadata can be very useful for improving classification accuracy. A. Blum & T. Mitchell [4] compared two approaches, one based on full-text, and one based on anchor words pointing to the target pages, and showed that anchor words alone were slightly less powerful than the full-text alone, and that the combination of the two was best. E. Glover et al. [11] claimed that text in citing documents close to the citation often had greater discriminative and descriptive power than text in target documents. Similarly, A. Attardi, A. Gulli & F. Sebastiani [3] also used information from the context where a URL that refers to that document appears and got encouraging results. J. Fürnkranz [9] used portions of texts from all pages that point to the target page: the anchor text, the headings that structurally precede it, the text of the paragraph in which it occurs, and a set of (automatically extracted) linguistic phrases that capture syntactic role of the anchor text in the paragraph; headings and anchor text proved to be most useful.

On the other hand, R. Ghani, S. Slattery & Y. Yang [10] claim that including words from linked neighborhoods should be done carefully since the neighborhoods could be rather "noisy". Different data collections contain Web pages of various characteristics. If certain characteristics are common to the majority of Web pages in the collection, an appropriate approach taking advantage of those could be applied, but if the Web pages are very heterogeneous, it is difficult to take advantage of any of the Web-specific characteristics (cf. [22], [8], [18]).

2.2 Evaluation Challenge

The problem of deriving the correct interpretation of a document's subject matter has been much discussed in the library science and related literature. It has been reported that different people, whether users or subject indexers, would assign different subject terms or classes to the same document. Studies on inter-indexer and intra-indexer consistency report generally low indexer consistency ([15], p. 99-101). There are two main factors that seem to affect it: 1) higher exhaustivity and specificity of subject indexing both lead to lower consistency (indexers choose the same first term for the major subject of the document, but the consistency decreases as they choose more classes or terms); 2) the bigger the vocabulary, or, the more choices the indexers have, the less likely they will choose the same classes or terms (ibid.).

In this study we start from the assumption that manual classes in our data collection are correct, and compare results of automated classification against them. The classification system used in the study is Engineering Information (Ei), which is rather big (around 800 classes) and deep (five hierarchical levels), allowing many different choices. Without a thorough qualitative analysis of automatically assigned classes we cannot be sure if the classes assigned by the algorithm, which were not manually assigned, are actually wrong.

2.3 Description of the Algorithm

This study is based on an automated classification approach [2] that has been developed within the DESIRE project [6] to produce "All" Engineering [1], an experimental module of the manually created subject gateway Engineering Electronic Library (EELS) [7] (no longer maintained).

The algorithm classifies Web pages 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. Each time a match is found, the document is assigned the corresponding class, which is awarded a relevance score, based on which term is matched (single word, phrase, Boolean), the type of class matched (main or optional) (*weight[term]*), and the part of the Web page in which the match is found (*weight[loc]*). A match of a phrase (a number of words in exact order) or a Boolean expression (all terms must be present but in any order) is made more discriminating than a match of a single word; a main class is made more important than an optional class (in the Ei thesaurus, main class (code) is the class to use for the term, while optional class (code) is to be used under certain circumstances). A list of suggested classes and corresponding relevance scores (S) is produced using the following algorithm:

$$S = \sum_{locs} \left(\sum_{terms} \left(freq[loc_j][term_i] * weight[term_i] * weight[loc_j] \right) \right) . \quad (1)$$

Only classes with scores above a pre-defined cut-off value (cf. section 4.5) are selected as *the* classes for the document. Having experimented with different ap-

proaches for stemming and stop-word removal, the best results were gained when an expanded stop-word list was used, and stemming was not applied. For more information on the algorithm, see [2] and [13].

3 Methodology

3.1 Data Collection

The data collection used in the study comprises a selection of Web pages from the EELS subject gateway [7]. EELS Web pages have been selected and classified by librarians for end users of the gateway.

For the study, only pages in English were kept, the reason being that Ei captions and descriptors are in English. Also, some other pages were removed because they contained very little or no text. (The problem of pages containing hardly any text could be dealt with in the future, by propagating the class obtained for their subordinate pages.) The final data collection consisted of 1003 Web pages in the field of engineering.

The data were organized in a relational database. Each document in the database was assigned Ei classes derived from the following elements:

- title (Title);
- headings (Headings);
- metadata (Metadata); and,
- page's main text (Text).

Each class was automatically assigned a score indicating the degree of certainty that it is the correct one. Every document also had manually assigned Ei classes (Manual), against which the automatically assigned classes were compared.

3.2 Methods for Evaluation and Deriving Significance Indicators

Various measures have been used to evaluate different aspects of automated classification performance [21]. Effectiveness, the degree to which correct classification decisions have been made, is often evaluated using performance measures from information retrieval, such as precision and recall, and F1 measure being the harmonic mean of the two. Solutions have also been proposed to measure partial overlap, i.e. the *degree* of agreement between correct and automatically assigned classes (see, for example, [5]).

In this study, three methods have been used for evaluating and deriving the significance of different Web-page elements:

1. total and partial precision, recall, and F1 measures (using macroaveraging);
2. semantic distance; and,
3. multiple regression.

1. The Ei classification system has a solid hierarchical structure, allowing for a rather credible test on partial overlap. Three different levels of overlap were tested: total overlap; partial overlap of the first three digits, e.g. “932.1.” and “932.2.” are considered the same; and, partial overlap of the first two digits, e.g. “932” and “933” are considered the same. Partial overlap of the first four digits has not been conducted because there were few classes of five-digit length in the data collection.

2. In the literature, different similarity measures have been used for hypermedia navigation and retrieval (see, for example, [20]). Semantic distance, a numerical value representing the difference in meaning between two concepts or terms, is one of them. There are different ways in which to calculate it. For example, the measure of clicking distance in a directory-browsing tree can be used. We used the hierarchical structure of the Ei classification system as the means of obtaining the following (rather arbitrary) measures of semantic distance between any two classes:

- 4, when the classes differ already in the first digit (e.g. 601 vs. 901);
- 2, when the classes differ already in the second digit (e.g. 932 vs. 901);
- 1, when the classes differ in the third digit (e.g. 674.1 vs. 673.1); and
- 0.5, when the classes differ in the fourth digit (e.g. 674.1 vs. 674.2).

Those values reflect how the hierarchical system is structured; e.g. we say that class 6 and class 7 are more distant from each other than classes 63 and 64, which are in turn more distant in meaning than 635.1 and 635.2.

Calculations were conducted using the average distance between manually and automatically assigned classes. For each document, average distances were calculated for each of the four elements, and then the values were averaged for all the documents. When there was more than one manually assigned class per document, the semantic distance was measured between an automatically assigned class and that manually assigned class which was most similar to the automatically assigned one.

3. Multiple regression was used in a rather simplified way: scores assigned based on individual elements of a Web page were taken as independent variables, while the final score represented the dependent variable. The dependent variable was set to either 1000 or 0, corresponding to a correct or an incorrect class respectively.

4 Significance Indicators

4.1 General

In Table 1 basic classification characteristics and tendencies of our data collection are given. All the documents (1003) have at least one, and no more than six manually assigned classes, the majority having up to three classes. Manual assignment of classes was based on collection-specific classification rules.

Concerning automatically assigned classes based on different parts of a page, not all the pages have classes based on all of them. Classes based on text are assigned to the majority of documents, while those based on metadata to the least number of documents. Based on only title, headings, or metadata, less than 50% of the documents would get classified at all. On the average, per every document there are two

manually assigned classes, two classes based on title, four based on headings, nine based on metadata, and some 18 classes based on text.

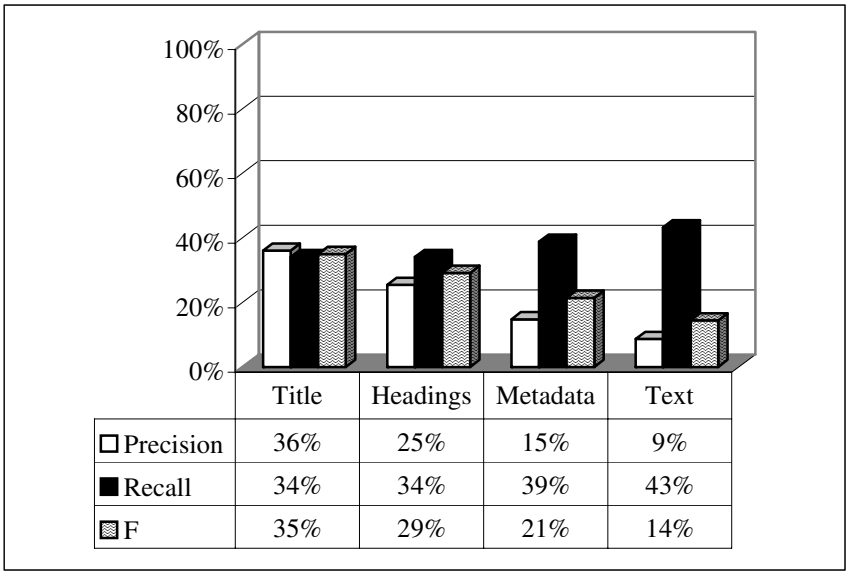
In the whole collection there are 753 different classes assigned, either manually or automatically. The largest variety comes from the group of classes assigned based on text (675), which is more than twice as many as manually assigned (305).

Table 1. Distribution of classes in the data collection. First data row shows how many documents have been classified, second row how many classes have been assigned in the whole of the data collection, and the last row how many different individual classes, out of some 800 possible, have been assigned.

	Manual	Title	Headings	Metadata	Text
Number of classified doc.	1003	411	391	260	964
In the data collection	1943	827	1504	2227	17089
Different classes	305	174	329	406	675

4.2 Precision and Recall

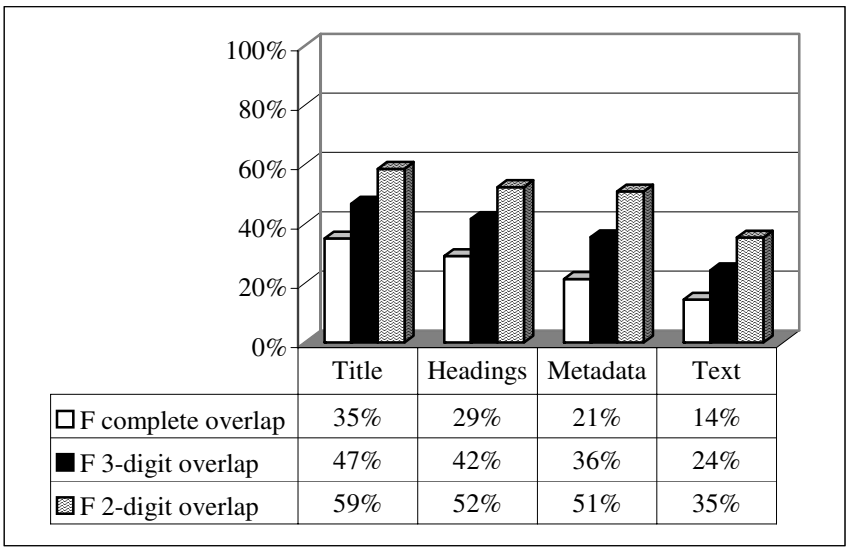
Fig. 1. shows the degree of automated classification accuracy when words are taken solely from the four different parts of the Web page. While title tends to yield best precision, which is 27% more than the worst element (text), text gives the best recall, but only 9% more than the worst element (title). Precision and recall are averaged using the F1 measure, according to which title performs the best (35%), closely followed by headings (29%), metadata (21%) and text (15%).



	Title	Headings	Metadata	Text
Precision	36%	25%	15%	9%
Recall	34%	34%	39%	43%
F	35%	29%	21%	14%

Fig. 1. Precision, recall and F1 measure

Partial Precision and Recall. When testing the algorithm performance for partial overlap (Fig. 2.), precision and recall for all parts of a Web page give much better results (title in 2-digit overlap achieves 59%). The ratio between their performance for both two- and three-digit overlap is the same as for total overlap: title performs the best, followed by headings, metadata and text.



	Title	Headings	Metadata	Text
F complete overlap	35%	29%	21%	14%
F 3-digit overlap	47%	42%	36%	24%
F 2-digit overlap	59%	52%	51%	35%

Fig. 2. F1-measure values for total overlap, 3-digit and 2-digit overlap

4.3 Semantic Distance

Using the semantic distance method, the calculations (Table 2) show that automatically assigned classes are on the average wrong in the third and second digits. Just like precision and recall results for partial overlap (cf. section 4.2), best results (smallest semantic distances) are achieved by title, followed by headings, metadata and text.

Table 2. Semantic distance

	Title	Headings	Metadata	Text
Mean distance	1,3	1,7	1,8	2,2

4.4 Deriving Significance Indicators

As we have seen in section 4.1, not every document has all the four elements containing sufficient terms for automated classification. Thus, in order to get documents classified, we need to use a combination of them. How to best combine them has been experimented with in this section, by applying results gained in evaluation using the F1 measure, semantic distance, and multiple regression.

The symbols used in formulae of this section are:

- S – final score for the automatically assigned class;
- STi – score for the automatically assigned class based on words in Title;
- SH – score for the automatically assigned class based on words in Headings;
- SM – score for the automatically assigned class based on words in Metadata; and,
- STe – score for the automatically assigned class based on words in Text.

The baseline, in which all the elements have equal significance, is represented with the following formula:

$$S = STi + SH + SM + STe . \quad (2)$$

Based on evaluation results, the following co-efficients, representing significance indicators, have been derived (the co-efficients were normalized by reducing the smallest co-efficient to one and by rounding others to integer values):

I. Based on total overlap and F1 measure values:

$$S = 2*STi + 2*SH + SM + STe . \quad (3)$$

These co-efficients have been derived by simply taking the F1 measure values of each of the algorithms (cf. Fig. 1). The same co-efficients have also been derived using partial overlap, the only difference being that the co-efficient for SM was two, both in two- and three-digit overlap.

II. Based on multiple regression, with scores not normalized for the number of words contained in title, headings, metadata, and text:

$$S = 86*STi + 5*SH + 6*SM + STe . \quad (4)$$

III. Based on multiple regression, with scores normalized for the number of words contained in title, headings, metadata, and text:

$$S = STi + SH + SM + 5*STe . \quad (\mathbf{5})$$

IV. On the basis of semantic distance results, the best significance indicator performs less than twice as well as the worst one, so all co-efficients are almost equal, as in (2).

4.5 Evaluation

Defining a Cut-Off. As described in section 2.3, each document is assigned a number of suggested classes and corresponding relevance scores. Only a few classes with best

scores, those above a certain cut-off value, are finally selected as *the* classes representing the document.

Different cut-offs, that would give best precision and recall results, were experimented with. Also, the number of documents that would be assigned at least one class, and the number of classes that would be assigned per document, were taken into consideration. Best results were achieved when the final classes selected were those with scores that contained at least 5% of all the scores assigned to all the classes, or, if such a class hadn't existed, the class with the top score was selected. In this case, F1 was 27%, there were about 4000 classes assigned as final, and all documents were classified. This is the cut-off we used in the study.

Results. As seen from Table 3, the evaluation showed that different significance indicators make hardly any difference in terms of classification algorithm performance. Co-efficients in (3) and (5) are similar to the ones in the baseline (2), and, compared to the baseline (2), which performs 23% in F1, normalized multiple regression (5) performs worse by 1%, while the formula based on F1 measure (3) performs the same. The best result was achieved using non-normalized multiple regression (4), which performs by 3% better than the baseline. This formula gives big significance indicator to classes that were assigned based on the title.

Table 3. Results of applying different co-efficients as significance indicators

	Baseline (2)	F1 (3)	Regression (4)	Regression N. (5)
Precision	16%	17%	21%	16%
Recall	39%	39%	35%	38%
F1	23%	23%	26%	22%
Number of pages	1003	1003	1003	1003
Number of classes	5174	5063	4073	5147

5 Conclusion

The aim of this study was to determine the significance of different parts of a Web page for automated classification: title, headings, metadata, and main text. The significance indicators were derived using several different methods: (total and partial) precision and recall, semantic distance, and multiple regression. The study showed that using *all* the structural elements and metadata is necessary since not all of them occur on every page. However, the exact way of combining the significance indicators turned out not to be highly important: the best combination of significance indicators is only 3% better than the baseline.

Reasons for such results need to be further investigated. One could guess that this is due to the fact that the Web pages in our data collection were rather heterogeneous; on the other hand, they were selected by librarians for end users of an operational service, and as such they might indicate what such Web-page collections are like. Apart from heterogeneity, the problem could be that metadata were abused, and that

certain tags were misused (e.g. instead of using appropriate tags for making text bold, one used a headings tag, which has the same effect on the screen).

Concerning evaluation of automated classification in general, further research is needed to determine the true value of the classification results. To that purpose information specialists and users could be involved, to compare their judgments as to which classes are correctly assigned. Also, in order to put the evaluation of classification into a broader context, a user study based on different information-seeking tasks would be valuable.

Other related issues of further interest include:

- determining significance of other elements, such as anchor text, location at the beginning of the document versus location at the end, etc.;
- comparing the results with new versions of the Web pages in the collection, e.g. maybe the quality of titles improves with time, and structural tags or metadata get less misused etc.; and,
- experimenting with other Web page collections.

Acknowledgements

The research was funded by ALVIS, an EU Sixth Framework Programme, Information Society Technologies (IST-1-002068-STP), and The Swedish Agency for Innovation Systems (P22504-1 A).

References

1. "All" Engineering Resources on the Internet: A Companion Service to EELS. Available: http://eels.lub.lu.se/ae/ (2003)
2. Ardö, A., Koch, T.: Automatic Classification Applied to the Full-Text Internet Documents in a Robot-Generated Subject Index. In: Online Information 99, Proceedings of the 23rd International Online Information Meeting, London. (1999) 239-246
3. Attardi, G., Gullì, A., Sebastiani, F.: Automatic Web Page Categorization by Link and Context Analysis. In: Hutchison, C., Lanzarone, G. (eds.): Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence. (1999) 105-119
4. Blum, A., Mitchell, T.: Combining Labeled and Unlabeled Data with Co-training. In: Annual Workshop on Computational Learning Theory, Proceedings of the Eleventh Annual Conference on Computational Learning Theory. (1998) 92-100
5. Ceci, M., Malerba, D.: Hierarchical Classification of HTML Documents with WebClassII. In: ECIR. (2003) 57-72
6. DESIRE : Development of a European Service for Information on Research and Education. Available: http://www.desire.org/ (2000)
7. Engineering Electronic Library. Available: http://eels.lub.lu.se/ (2003)
8. Fisher, M., Everson R.: When are Links Useful?: Experiments in Text Classification. In: Proceedings of ECIR-03, 25th European Conference on Information Retrieval, Pisa, IT (2003) 41-56
9. Fürnkranz, J.: Hyperlink Ensembles: A Case Study in Hypertext Classification. Information Fusion 3, 4 (2002) 299-312

10. Ghani, R., Slattery, S., Yang, Y.: Hypertext Categorization Using Hyperlink Patterns and Metadata. In: Proceedings of ICML-01, 18th International Conference on Machine Learning. (2001), 178-185
11. Glover, E.J. et al.: Using Web structure for Classifying and Describing Web Pages. In: Proceedings of the Eleventh International Conference on World Wide Web Honolulu, Hawaii, USA. (2002) 562–569
12. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys 3, 31 (1999) 264-323
13. Koch, T., Ardö, A.: Automatic Classification of Full-Text HTML-Documents from One Specific Subject Area. EU Project DESIRE II D3.6a, Working Paper 2. Available: http://www.it.lth.se/knowlib/DESIRE36a-WP2.html. (2000)
14. Kolcz, A., Prabakarmurthi, V., Kalita, J., and Alspector, J.: Summarization as Feature Selection for Text Categorization. In: Proceedings of the Tenth International Information and Knowledge Management (CIKM-01). (2001) 365-370
15. Olson, H.A., Boll, J.J.: Subject Analysis in Online Catalogs. 2nd ed. Libraries Unlimited, Englewood, Colorado (2001)
16. Pierre, J.: On the Automated Classification of Web sites. In: Linköping Electronic Articles in Computer and Information Science 001 (6) (2001). Available: http://www.ep.liu.se/ea/cis/2001/001/
17. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 1, 34 (2002) 1–47
18. Slattery, S., Craven, M.: Discovering Test Set Regularities in Relational Domains. In: Proceedings of ICML-00, 17th International Conference on Machine Learning. (2000), 895-902
19. Svenonius, E.: The Intellectual Foundations of Information Organization. MIT Press, Cambridge, MA (2000)
20. Tudhope, D., Taylor C.: Navigation via Similarity: Automatic Linking Based on Semantic Closeness. Information Processing and Management, 33(2) (1997) 233-242
21. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval 1/2, 1 (1999) 67-88
22. Yang, Y., Slattery, S., Ghani, R.: A Study of Approaches to Hypertext Categorization. Journal of Intelligent Information Systems. 2/3, 8 (2002) 219-241