Automated subject classification of textual web documents

Koraljka Golub
Department of Information Technology, Lund University, Lund, Sweden

Abstract
Purpose – To provide an integrated perspective to similarities and differences between approaches to automated classification in different research communities (machine learning, information retrieval and library science), and point to problems with the approaches and automated classification as such.
Design/methodology/approach – A range of works dealing with automated classification of full-text web documents are discussed. Explorations of individual approaches are given in the following sections: special features (description, differences, evaluation), application and characteristics of web pages.
Findings – Provides major similarities and differences between the three approaches: document pre-processing and utilization of web-specific document characteristics is common to all the approaches; major differences are in applied algorithms, employment or not of the vector space model and of controlled vocabularies. Problems of automated classification are recognized.
Research limitations/implications – The paper does not attempt to provide an exhaustive bibliography of related resources.
Practical implications – As an integrated overview of approaches from different research communities with application examples, it is very useful for students in library and information science and computer science, as well as for practitioners. Researchers from one community have the information on how similar tasks are conducted in different communities.
Originality/value – To the author's knowledge, no review paper on automated text classification attempted to discuss more than one community's approach from an integrated perspective.
Keywords Automation, Classification, Internet, Document management, Controlled languages
Paper type Literature review

1. Introduction

Classification is, to the purpose of this paper, defined as:
... the multistage process of deciding on a property or characteristic of interest, distinguishing things or objects that possess that property from those which lack it, and grouping things or objects that have the property or characteristic in common into a class. Other essential aspects of classification are establishing relationships among classes and making distinctions within classes to arrive at subclasses and finer divisions (Chan, 1994, p. 259).
Automated subject classification (in further text: automated classification) denotes machine-based organization of related information objects into topically related groups. In this process human intellectual processes are replaced by, for example, statistical and computational linguistics techniques. In the literature on automated classification, the terms automatic and automated are both used. Here the term automated is chosen because it more directly implies that the process is machine-based. Automated classification has been a challenging research issue for several

decades now. Major motivation has been the high cost of manual classification. Interest has grown rapidly since 1997, when search engines could not do with just text retrieval techniques, because the number of available documents grew exponentially. Owing to the ever-increasing number of documents, there is a danger that recognized objectives of bibliographic systems would get left behind; automated means could be a solution to preserve them (Svenonius, 2000, pp. 20-1, 30). Automated classification of text finds its use in a wide variety of applications, such as: organizing documents into subject categories for topical browsing, including grouping search results by subject; topical harvesting; personalized routing of news articles; filtering of unwanted content for internet browsers; and many others (Sebastiani, 2002; Jain et al., 1999).

In the narrower focus of this paper is automated classification of textual web documents into subject categories for browsing. Web documents 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 misused, structural tags can also be misused, and titles can be general ("home page" "untitled document"). Browsing in this paper refers to seeking for documents via a hierarchical structure of subject classes into which the documents had been classified. Research has shown that people find browsing useful in a number of information-seeking situations, such as: when not looking for a specific item, when one is inexperienced in searching (Koch and Zettergren, 1999), or unfamiliar with the subject in question and its terminology or structure (Schwartz, 2001, p. 76).

In the literature, terms such as classification, categorization and clustering are used to represent different approaches. In their broadest sense these terms could be considered synonymous, which is probably one of the reasons why they are interchangeably used in the literature, even within the same research communities. For example, Hartigan (1996, p. 2) says: "The term cluster analysis is used most commonly to describe the work in this book, but I much prefer the term classification..." Or: ".. . classification or categorization is the task of assigning objects from a universe to two or more classes or categories" (Manning and Schu¨ tze, 1999, p. 575).

In this paper terms text categorization and document clustering are chosen because they tend to be the prevalent terms in the literature of the corresponding communities. Document classification and mixed approach are used in order to consistently distinguish between the four approaches. Descriptions of the approaches are given below:

(1) Text categorization. It is a machine-learning approach, in which also information retrieval methods are applied. It consists of three main parts: categorizing a number of documents to pre-defined categories, learning the characteristics of those documents, and categorizing new documents. In the machine-learning terminology, text categorization is known as supervised learning, since the process is "supervised" by learning categories' characteristics from manually categorized documents.

(2) Document clustering. It is an information-retrieval approach. Unlike text categorization, it does not involve pre-defined categories or training documents and is thus called unsupervised. In this approach the clusters and, to a limited degree, relationships between clusters are derived automatically from the documents to be clustered, and the documents are subsequently assigned to those clusters.

(3) Document classification. In this paper it stands for a library science approach. It involves an intellectually created controlled vocabulary (such as classification schemes), into classes of which documents are classifi d. Controlled vocabularies have been developed and used in libraries and in indexing and abstracting services, some since the end of the 19th century.

(4) Mixed approach. Sometimes methods from text categorization or document clustering are

used together with controlled vocabularies. In the paper such an approach is referred to as a mixed approach.
To the author's knowledge no review paper on automated text classification attempted to discuss more than one community's approach. Individual approaches of text categorization (document) clustering and document classification have been analysed by Sebastiani (2002), Jain et al. (1999) and Toth (2002), respectively.
This paper deals with all the approaches, from an integrated perspective. It is not aimed at detailed descriptions of approaches, since they are given in the above-mentioned reviews. Nor does it attempt to be comprehensive and all-inclusive. It aims to point to similarities or differences as well as problems with the existing approaches. In what aspects and to what degree are today's approaches to automated classification comparable? To what degree can the process of subject classification really be automated, with the tools available today? What are the remaining challenges? These are the questions touched upon in the paper.
The paper is laid out as follows: explorations of individual approaches as to their special features (description, differences, evaluation), application and employment of characteristics of web pages are given in the second section (approaches to automated classification), followed by a discussion (third section).

2. Approaches to automated classification

2.1 Text categorization

2.1.1 Special features.

2.1.1.1 Description of features. Text categorization is a machine-learning approach, which has also adopted some features from information retrieval. The process of text categorization consists of three main parts:
(1) The first part involves manual categorization of a number of documents to pre-defined categories. Each document is represented by a vector of terms. (The vector space model comes from information retrieval). These documents are called training documents because, based on those documents, characteristics of categories they belong to are learnt.
(2) By learning the characteristics of training documents, for each category a program called classifier is constructed. After the classifiers have been created, and before automated categorization of new documents takes place, classifiers are tested with a set of so-called test documents, which were not used in the first step.
(3) The third part consists of applying the classifier to new documents.

In the literature, text categorization is known as supervised learning, since the process is "supervised" by learning from manually pre-categorized documents. As opposed to text categorization, clustering is known as an unsupervised approach, because it does not involve manually pre-clustered documents to learn from. Nonetheless, due to the fact that manual pre-categorization is rather expensive, semi-supervised approaches, which diminish the need for a large number of training documents, have also been implemented (Blum and Mitchell, 1998; Liere and Tadepalli, 1998; McCallum et al., 2000).
2.1.1.2 Differences within the approach. A major difference among text categorization approaches is in how classifiers are built. They can be based on Bayesian probabilistic learning, decision tree learning, artificial neural networks, genetic algorithms or instance-based learning – for explanation of those, see, for example, Mitchell (1997). There have also been attempts of classifier committees (or metaclassifiers), in which results of a number of different classifiers are combined to decide on a category (e.g. Liere and Tadepalli, 1998). One also needs to mention that not all algorithms used in

text categorization are based on machine learning. For example, Rocchio (1971) is actually an information retrieval classifier and WORD (Yang, 1999) is a non-learning algorithm, invented to enable comparison of learning classifiers' categorization accuracy. Comparisons of learning algorithms can be found in Schu¨ tze et al. (1995), Li and Jain (1998), Yang (1999) or Sebastiani (2002).
Another difference within the text categorization approach is in the document pre-processing and indexing part, where documents are represented as vectors of term weights. Computing the term weights can be based on a variety of heuristic principles. Different terms can be extracted for vector representation (single words, phrases, stemmed words, etc.), also based on different principles; characteristics of web documents, such as mark-up for emphasized terms and links to other documents, are often experimented with (Go¨vert et al., 1999). The number of terms per document needs to be reduced not only for indexing the document with most representative terms, but also for computing reasons. This is called dimensionality reduction of the term space. Dimensionality reduction methods could include removal of non-informative terms (not only stop words); also, taking only parts of the web document, its snippet or summary (Mladenic and Grobelnik, 2003), has been explored. For an example of a complex document representation approach, a word clustering one, see Bekkerman et al. (2003); for another example, based on latent semantic analysis, see Cai and Hofmann (2003).
Several researches have explored how hierarchical structure of categories into which documents are to be categorized could influence the categorization performance. Koller and Sahami (1997) used a Bayesian classifier at each node of the classification hierarchy and employed a feature selection method to find a set of discriminating features (i.e. words) for each node. They showed that, in comparison to a flat approach, using hierarchical structure could improve classification performance. Similar improvements were reported by McCallum et al. (1998), Dumais and Chen (2000) and Ruiz and Srinivasan (1999).

2.1.1.3 Evaluation methods. Various measures are used to evaluate different aspects of text categorization performance (Yang, 1999). Effectiveness, the degree to which correct categorization decisions have been made, is often evaluated using performance measures from information retrieval, such as precision (correct positives/predicted positives) and recall (correct positives/actual positives). Efficiency can also be evaluated, in terms of computing time spent on different parts of the process. There are other evaluation measures, and new are being developed such as those that take into account degrees to which a document was wrongly categorized (Dumais et al., 2002; Sun et al., 2001). For more on evaluation measures in text categorization, see Sebastiani (2002, p. 32-9). Evaluation in text categorization normally does not involve subject experts or users.
Yang (1999) claims that the most serious problem in text categorization evaluations is the lack of standard data collections and shows how different versions of the same collection have a strong impact on the performance, and other versions do not. Some of the data collections used by the text categorization community are: Reuters-21578 (2004), which contains newswire stories classified under categories related to economics; OHSUMED (Hersh, 1994), containing abstracts from medical journals categorized under Medical Subject Headings (MeSH); the US Patent database in which patents are categorized into the US Patent Classification System; 20 Newsgroups DataSet (1998) containing about 20,000 postings to 20 different Usenet newsgroups. For web documents there is WebKB (2001), Cora (McCallum et al., 1999), and samples from directories of web documents such as Yahoo! (Yahoo!, 2005). All these collections have a different number of categories and hierarchical levels. There seems to be a tendency to conduct experiments on a relatively small number of categories with few hierarchical levels, which is usually not suitable for subject browsing tasks.

2.1.2 Characteristics of web pages. A number of issues related to categorization of textual web documents have been dealt with in the literature. Hypertext-specific characteristics such as hyperlinks, HTML tags and metadata have all been explored. Yang et al. (2002) have defined five hypertext regularities of web document collections, which need to be recognized in order to chose an appropriate text categorization approach:

(1) no hypertext regularity; in which case standard classifiers for text are used;

(2) encyclopaedia regularity, when documents with a certain category label only link to documents with the same category label, in which case the text of each document could be augmented with the text of its neighbours;

(3) co-referencing regularity, when neighbouring documents have a common topic; in which case the text of each document can be augmented with the text of its neighbours, but text from the neighbours should be marked (e.g. prefixed with a tag);

(4) preclassified regularity, when a single document contains hyperlinks to documents with the same topic, in which case it is sufficient to represent each page with names of the pages it links with; and

(5) metadata regularity, when there are either external sources of metadata for the documents on the web, in which case we extract the metadata and look for features that relate documents being categorized, or metadata are contained within the META, ALT and TITLE tags.

Several other papers discuss characteristics of document collections to be categorized. Chakrabarti et al. (1998b) showed that including documents that cite, or are cited by the document being categorized, as if they were local terms, performed worse than when those documents were not considered. They achieved improved results applying a more complex approach with refining the class distribution of the document being classified, in which both the local text of a document and the distribution of the estimated classes of other documents in its neighbourhood, were used. Slattery and Craven (2000) showed how discovering regularities, such as words occurring on target pages and on other pages related by hyperlinks, in both training and test document sets could improve categorization accuracy. Fisher and Everson (2003) found out that link information could be useful if the document collection had a sufficiently high link density and links were of sufficiently high quality. They introduced a frequency-based method for selecting the most useful citations from a document collection.

Blum and Mitchell (1998) compared two approaches, one based on full-text, and the other based on anchor words, and found out that anchor words alone were slightly less powerful than the full-text alone, and that the combination of the two was best. Glover et al. (2002) reported that the text in citing documents close to the citation often has greater discriminative and descriptive power than the text in the target document. Similarly, Attardi et al. (1999) used information from the context where a URL that refers to that document appears and got encouraging results. Fu¨ rnkranz (1999) included words that occurred in nearby headings and in the same paragraph as anchor-text, which yielded better results than using the full-text alone. In his later study Fu¨ rnkranz (2002) 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 linguistic phrases that capture syntactic role of the anchor text in this paragraph. Headings and anchor text seemed to be most useful.

In regards to metadata, Ghani et al. (2001) reported that metadata could be very useful for improving classification accuracy.

2.1.3 Application. Text categorization is the most frequently used approach to automated classification. While a large portion of research is aimed at improving algorithm performance, it has

been applied in operative information systems, such as Cora (McCallum et al., 2000), NorthernLight (Dumais et al., 2002, pp. 69-70) and the Thunderstone's Web Site Catalog (Thunderstone, 2005). However, detailed information about approaches used in commercial directories is mostly not available, due to their proprietary nature (Pierre, 2001, p. 9). There are other examples of applying machine-learning techniques to web pages and categorizing them into browsable structures. Mladenic (1998) and Labrou and Finin (1999) used the Yahoo! Directory (Yahoo!, 2005). Pierre (2001) categorized web pages into industry categories, although he used only top-level categories of North American Industrial Classification System.
Apart from organizing web pages into categories, text categorization has been applied for categorizing web search engine results (Chen and Dumais, 2000; Sahami et al., 1998). It also finds its application in document filtering, word sense disambiguation, speech categorization, multimedia document categorization, language identification, text genre identification, and automated essay grading (Sebastiani, 2002, p. 5).

2.1.4 Summary. Text categorization is a machine-learning approach, with the vector-space model and evaluation measures borrowed from information retrieval. Characteristics of pre-defined categories are learnt from manually categorized documents. Within text categorization, differences occur in several aspects: algorithms, methods applied to represent documents as vectors of term weights, evaluation measures and data collections used.
The potential added value of web document characteristics, which have been compared and experimented with, are, for example, anchor words, headings words, text near the URL for the target document, inclusion of linked document's text as being local. When deciding which methods to use, one needs to determine which characteristics are common to the documents to be categorized; for example, augmenting the document to be classified with the text of its neighbours will yield good results only if the source and the neighbours are related enough.
Text categorization is the most widespread approach to automated classification, with a lot of experiments being conducted under controlled conditions. There seems to be a tendency to use a small number of categories with few hierarchical levels, which is usually not suitable for subject browsing tasks. Several examples of its application in operative information systems exist.

2.2 Document clustering
2.2.1 Special features.
2.2.1.1 Description of features. Document clustering is an information retrieval approach. As opposed to text categorization, it does not involve manually pre-categorized documents to learn from, and is thus known as an unsupervised approach.
The process of document clustering involves two main steps:
(1) Documents to be clustered are represented by vectors, which are then compared to each other using similarity measures. Like in text categorization, different principles can be applied at this stage to derive vectors (which words or terms to use, how to extract them, which weights to assign based on what, etc.). Also, different similarity measures can be used, the most frequent one probably being the cosine measure.
(2) In the following step, documents are grouped into clusters using clustering algorithms. Two different types of clusters can be constructed: partitional (or flat), and hierarchical.
Partitional algorithms determine all clusters at once. A usual example is K-means, in which first a k number of clusters is randomly generated; when new documents are assigned to the nearest centroid (centre of a cluster), centroids for clusters need to be re-computed.
In hierarchical clustering, a hierarchy of clusters is built. Often agglomerative algorithms are used:

first, each document is viewed as an individual cluster; then, the algorithm finds the most similar pair of clusters and merges them. Similarity between documents can be calculated in a number of ways. For example, it can be defined as the maximum similarity between any two individuals, one from each of the two groups (single-linkage), as the minimum similarity (complete-linkage), or as the average similarity (group-average linkage). For a review of different clustering algorithms, see Jain et al. (1999), Rasmussen (1992) and Fasulo (1999).
Another approach to document clustering is self-organizing maps (SOMs). SOMs are a data visualisation technique, based on unsupervised artificial neural networks, that transform high-dimensional data into (usually) two-dimensional representation of clusters. For a detailed overview of SOMs, see Kohonen (2001). There are several research examples of visualization for browsing using SOMs (Heuser et al., 1998; Poincot et al., 1998; Rauber and Merkl, 1999; Goren-Bar et al., 2000; Schweighofer et al., 2001; Yang et al., 2003; Dittenbach et al., 2004).

2.2.1.2 Differences within the approach. A major difference within the document clustering community is in algorithms (see above). While previous research showed that agglomerative algorithms performed better than partitional ones, some studies indicate the opposite. Steinbach et al. (2000) compared agglomerative hierarchical clustering and K-means clustering and showed that K-means is at least as good as agglomerative hierarchical clustering. Zhao and Karypis (2002) evaluated different partitional and agglomerative approaches and showed that partitional algorithms always lead to better clustering solutions than agglomerative algorithms. In addition, they presented a new type of clustering algorithms called constrained agglomerative algorithms that combined the features of both partitional and agglomerative algorithms. This solution gave better results than agglomerative or partitional algorithms alone. For a comparison of hierarchical clustering algorithms, and added value of some linguistics features, see Hatzivassiloglou et al. (2000). Different enhancements to algorithms have been proposed (Liu et al., 2002; Mandhani et al., 2003; Slonim et al., 2003).
Since, in document clustering (including SOMs) clusters and their labels are produced automatically, deriving the labels is a major research challenge. In an early example of automatically derived clusters (Garfield et al., 1975), which were based on citation patterns, labels were assigned manually. Today a common heuristic principle is to extract between five and ten of the most frequent terms in the centroid vector, then to drop stop-words and perform stemming, and choose the term which is most frequent in all documents of the cluster. A more complex approach to labelling is given by Glover et al. (2003). They used an algorithm to predict “parent, self, and child terms”; self terms were assigned as clusters’ labels, while parent and children terms were used to correctly position clusters in the cluster collection.
Another problem in document clustering is how to deal with large document collections. According to Jain et al. (1999, p. 316), only the K-means algorithm and SOMs, have been tested on large data sets. An example of an approach dealing with large data sets and high dimensional spaces was presented by Haveliwala et al. (2000), who developed a technique they managed to apply to 20 million URLs.
2.2.1.3 Evaluation methods. Similarly to text categorization, there are many evaluation measures (e.g. precision and recall), and evaluation normally does not include subject experts or users.
Data collections often used are fetched from TREC (2004). In the development stage is the INEX initiative project (INitiative for the Evaluation of XML Retrieval, 2004), within which a large data collection of XML documents, over 12,000 articles from IEEE publications from the period of 1995-2002, would be provided.

2.2.2 Characteristics of web pages. A number of researchers have explored the potential of hyperlinks in the document clustering process. Weiss et al. (1996) were assigning higher similarities to documents that have ancestors and descendants in common. Their preliminary results also illustrated that combining term and link information yields improved results. Wang and Kitsuregawa (2002) experimented with best ways of combining terms from web pages with words from in-link pages (pointing to the web page) and out-link pages (leading from the web page), and achieved improved results.

Other web-specific characteristics have been explored. Information about users' traversals in the category structure has been experimented with (Chen et al., 2002), as well as usage logs (Su et al., 2001). The hypothesis behind this approach is that the relevancy information is objectively reflected by the usage logs; for example, it is assumed that frequent visits by the same person to two seemingly unrelated documents indicate that they are closely related.

2.2.3 Application. Clustering is the unsupervised classification of objects, based on patterns (observations, data items, feature vectors) into groups or clusters (Jain et al., 1999, p. 264). It has been addressed in various disciplines for many different applications (Jain et al., 1999, p. 264); in information retrieval, documents are the ones that are grouped or clustered (hence the term document clustering).

Traditionally, document clustering has been applied to improve document retrieval (for a review, see Willet, 1988; for an example, see Tombros and van Rijsbergen, 2001). In this paper the emphasis is on automated generation of hierarchical clusters structure and subsequent assignment of documents to those clusters for browsing.

An early attempt to cluster a document collection into clusters for the purpose of browsing was Scatter/Gather (Cutting et al., 1992). Scatter/Gather would partition the collection into clusters of related documents, present summaries of the clusters to the user for selection, and when the user would select a cluster, the narrower clusters were presented; when the narrowest cluster would be reached, documents were enumerated. Another approach is presented by Merchkour et al. (1998). First the so-called source collection (an authoritative collection representative in the domain of interest of the users) would be clustered for the user to browse it, with the purpose of helping him/her with defining the query. Then the query would be submitted via a web search engine to the target collection, which is the world wide web. The results would be clustered into the same categories as in the source collection. Kim and Chan (2003) attempted to build a personalized hierarchy for an individual user, from a set of web pages the user visited, by clustering words from those pages. Other research has been conducted in automated construction of vocabularies for browsing (Chakrabarti et al., 1998a; Wacholder et al., 2001).

Another application of automated generation of hierarchical category structure and subsequent assignment of documents to those categories is organization of web search engine results (Clusty, 2004; MetaCrawler Web search, 2005; Zamir et al., 1997; Zamir and Etzioni, 1998; Palmer et al., 2001; Wang and Kitsuregawa, 2002).

2.2.4 Summary. Like in text categorization, in document clustering documents are first represented as vectors of term weights. Then they are compared for similarity, and grouped into partitional or hierarchical clusters using different algorithms. Characteristics of web documents similar to those from text categorization approach have been explored.

In evaluation, precision, recall and other measures are used, while end-users and subject experts are normally left out.
Unlike text categorization, document clustering does not require either training documents, or pre-existing categories into which the documents are to be grouped. The categories are created when groups are formed – thus, both the names of the groups and relationships between them are automatically derived. The derivation of names and relationships is the most challenging issue in document clustering.
Document clustering was traditionally used to improve information retrieval. Today it is better suited for clustering search-engine results than for organizing a collection of documents for browsing, because automatically derived cluster labels and relationships between the clusters are incorrect or inconsistent. Also, clusters change as new documents are added to the collection – such instability of browsing structure is not user-friendly either.

2.3 Document classification

2.3.1 Special features.

2.3.1.1 Description of features. Document classification is a library science approach. The tradition of automating the process of subject determination of a document and assigning it to a term from a controlled vocabulary partly has its roots in machine-aided indexing (MAI). MAI has been used to suggest controlled vocabulary terms to be assigned to a document.
The automated part of this approach differs from the previous two in that it is generally not based on either supervised or unsupervised learning. Neither do documents and classes get represented by vectors. In document classification, the algorithm typically compares terms extracted from the text to be classified, to terms from the controlled vocabulary (string-to-string matching). At the same time, this approach does share similarities with text categorization and document clustering: the pre-processing of documents to be classified includes stop-words removal; stemming can be conducted; words or phrases from the text of documents to be classified are extracted and weights are assigned to them based on different heuristics. Web-page characteristics have also been explored, although to a lesser degree.
The most important part of this approach is controlled vocabularies, most of which have been created and maintained for use in libraries and indexing and abstracting services, some of them for more than a century. These vocabularies have devices to “control” polysemy, synonymy, and homonymy of the natural language. They can have systematic hierarchies of concepts, and a variety of relationships defined between the concepts. There are different types of controlled vocabularies, such as classification schemes, thesauri and subject heading systems. With the world wide web, new types of vocabularies emerged within the computer science and the semantic web communities: ontologies and search-engine directories of web pages. All these vocabularies have distinct characteristics and are consequently better suited for some classification tasks and applications than others (Koch and Day, 1997; Koch and Zettergren, 1999; Vizine-Goetz, 1996). For example, subject heading systems normally do not have detailed hierarchies of terms (exception: medical subject headings), while classification schemes consist of hierarchically structured groups of classes. The latter are better suited for subject browsing. Also, different classification schemes have different characteristics of hierarchical levels. For subject browsing the following are important:

the bigger the collection, the more depth should the hierarchy contain; classes should contain more than just one or two documents (Schwartz, 2001, p. 48). On the other hand, subject heading systems and thesauri have traditionally been developed for subject indexing to describe topics of the document as specifically as possible. Since, both classification schemes and subject headings or

thesauri provide users with different aspects of subject information and different searching functions, their combined usage has been part of practice in indexing and abstracting services. Ontologies are usually designed for very specific subject areas and provide rich relationships between terms. Search-engine directories and other home-grown schemes on the web:
... even those with well-developed terminological policies such as Yahoo .. . suffer from a lack of understanding of principles of classification design and development. The larger the collection grows, the more confusing and overwhelming a poorly designed hierarchy becomes... (Schwartz, 2001, p. 76).
Although well-structured and developed, existing controlled vocabularies need to be improved for the new roles in the electronic environment. Adjustments should include:

- improved currency and capability for accommodating new terminology;
- flexibility and expandability – including possibilities for decomposing faceted notation for retrieval purposes;
- intelligibility, intuitiveness, and transparency – it should be easy to use, responsive to individual learning styles, able to adjust to the interests of users, and allow for custom views;
- universality – the scheme should be applicable for different types of collections and communities and should be able to be integrated with other subject languages; and
- authoritativeness – there should be a method of reaching consensus on terminology, structure, revision, and so on, but that consensus should include user communities ([10], pp. 77-8).

Some of the controlled vocabularies are already being adjusted, such as: AGROVOC, the agricultural thesaurus (Soergel et al., 2004), WebDewey, which is the Dewey Decimal Classification (DDC, 2005) adapted for the electronic environment, and California Environmental Resources thesaurus (CERES, 2003).

2.3.1.2 Differences within the approach. The differences occur in document pre-processing, which includes word or phrase extraction, stemming, etc. heuristic principles (such as weighting based on where the term/word occurs or occurrence frequency), linguistic methods, and controlled vocabulary applied.
The first major project aimed at automated classification of web pages based on a controlled vocabulary was the Nordic WAIS/World Wide Web Project (1995), which took place at Lund University Library and National Technological Library of Denmark (Ardo¨ et al., 1994; Koch, 1994). In this project automated classification of the world wide web and Wide Area Information Server (WAIS) databases using Universal Decimal Classification (UDC) was experimented with. A WAIS subject tree was built based on two top levels of UDC, i.e. 51 classes. The process involved the following steps: words from different parts of database descriptions were extracted, and weighted based on which part of the description they belonged to; by comparing the extracted words with UDC's vocabulary a ranked list of suggested classifications was generated. The project started in 1993, and ended in 1996, when WAIS databases came out of fashion.
GERHARD is a robot-generated web index of web documents in Germany (GERHARD, 1999, 1998; Mo¨ller et al., 1999). It is based on a multilingual version of UDC in English, German and French, adapted by the Swiss Federal Institute of Technology Zurich (Eidgeno¨ssische Technische Hochschule Zu¨ rich – ETHZ). GERHARD's approach included advanced linguistic analysis: from captions, stop words were removed, each word was morphologically analysed and reduced to stem; from web pages stop words were also removed and prefixes were cut off. After the linguistic analysis, phrases were extracted from the web pages and matched against the captions. The resulting set of UDC notations was ranked and weighted statistically, according to frequencies and document structure.
Online Computer Library Center's (OCLC) project Scorpion (2004) built tools for automated subject

recognition, using DDC. The main idea was to treat a document to be indexed as a query against the DDC knowledge base. The results of the "search" were treated as subjects of the document. Larson (1992) used this idea earlier, for books. In Scorpion, clustering was also used, for refining the result set and for further grouping of documents falling in the same DDC class (Subramanian and Shafer, 1998). The System for Manipulating and Retrieving Text (SMART) weighting scheme was used, in which term weights were calculated based on several parameters: the number of times that the term occurred in a record; how important the term was to the entire collection based on the number of records in which it occurred; and, the normalization value, which is the cosine normalization that computes the angle between vector representations of a record and a query. Different combinations of these elements have been experimented with. Another OCLC project, WordSmith (Godby and Reighart, 1998), was to develop software to extract significant noun phrases from a document. The idea behind it was that the precision of automated classification could be improved if the input to the classifier were represented as a list of the most significant noun phrases, instead as the complete text of the raw document. However, it showed that there were no significant differences. OCLC currently works on releasing FAST (2004), based on the Library of Congress Subject Headings (LCSH), which are modified into a post-coordinated faceted vocabulary. The eight facets to be implemented are: topical, geographic (place), personal name, corporate name, form (type, genre), chronological (time, period), title and meeting place. FAST could also serve as a knowledge base for automated classification, like the DDC database did in Scorpion (FAST, 2003).
Wolverhampton Web Library (WWLib) is a manually maintained library catalogue of British web resources, within which experiments on automating its processes were conducted (Wallis and Burden, 1995; Jenkins et al., 1998). Original classifier from 1995 was based on comparing text from each document to DDC captions. In 1998 each classmark in the DDC captions file was enriched with additional keywords and synonyms. Keywords extracted from the document were weighted on the basis of their position in the document. The classifier began by matching documents against class representatives of top ten DDC classes and then proceeded down through the hierarchy to those subclasses that had a significant measure of similarity (Dice's coefficient) with the document.

"All" Engineering (EELS, 2003) is a robot-generated web index of about 300,000 web documents, developed within DESIRE (DESIRE project, 1999; DESIRE, 2000), as an experimental module of the manually created subject gateway Engineering Electronic Library (EELS) (Koch and Ardo¨ 2000; Engineering Electronic Library, 2003). Engineering Index (Ei) thesaurus was used; in this thesaurus, terms are enriched with their mappings to Ei classes. Both Ei captions and thesaurus terms were matched against the extracted title, metadata, headings and plain text of a full-text document from the world wide web. Weighting was based on term complexity and type of classification, location and frequency. Each pair of term-class codes was assigned a weight depending on the type of term (Boolean, phrase, single word), and the type of class code (main code, the class to be used for the term, or optional code, the class to be used under certain circumstances); a match of a Boolean expression or a phrase was made more discriminating than a match of a single word; a main code was made more important than an optional code. Having experimented with different approaches for stemming and stop-word removal, the best results were gained when an expanded stop-word list was used, and stemming was not applied. The DESIRE project proved the importance of applying a good controlled vocabulary in achieving the classification accuracy: 60 per cent of documents were correctly classified, using only a very simple algorithm based on a limited set of heuristics and simple weighting. Another robot-generated web index, Engine-e (2004), used a slightly modified automated classification approach to the one developed in "All" Engineering (Lindholm et al., 2003). Engine-e provided subject browsing of engineering documents based on Ei terms, with six broader categories

as starting points.
The project Bilingual Automatic Parallel Indexing and Classification (BINDEX, 2001; Nu¨ bel et al., 2002) was aimed at indexing and classifying abstracts from engineering in English and German, using English INSPEC thesaurus and INSPEC classification, FIZ Technik's bilingual thesaurus, "Engineering and Management" and the Classifi on Scheme, "Fachordnung Technik 1997". They performed morpho-syntactic analysis of a document, which consisted of identification of single and multiple-word terms, tagging and lemmatization, and homograph resolution. The extracted keywords were checked against the INSPEC thesaurus and the German part of "Engineering and Management" and classification codes were derived. Keywords which were not in the thesaurus were assigned as free indexing terms.
2.3.1.3 Evaluation methods. Measures such as precision and recall have been used. This approach differs from the other two approaches in that evaluation of document classification tends to also involve subject experts or intended users (Koch and Ardo¨ 2000), which is in line with traditional library science evaluations.
Examples of data collections that have been used are harvested web documents (GERHARD, "All" Engineering), and bibliographic records of internet resources (Scorpion).
2.3.2 Summary. Document classification is a library science approach. It differs from text categorization and document clustering in that well-developed controlled vocabularies are employed, whereas vector space model and algorithms based on vector calculations are generally not used. Instead, selected terms from documents to be classified are compared against terms in the chosen controlled vocabulary, whereby often computational linguistic techniques are employed.
In evaluation, performance measures from information retrieval are used, and, unlike the other two approaches, subject experts or users tend to be involved.
In the focus of research are mainly publicly available operative information systems that provide browsing access to their document collections.

2.4 Mixed approach
Mixed approach is the term used here to refer to a machine-learning or an information-retrieval approach, in which also controlled vocabularies that have been traditionally used in libraries and indexing and abstracting services are used. There do not seem to be many examples of this approach. Frank and Paynter (2004) applied machine-learning techniques to assign Library of Congress Classification (LCC) notations to resources that already have an LCSH term assigned. Their solution has been applied to INFOMINE (subject gateway for scholarly resources, http://infomine. ucr.edu/), where it is used to support hierarchical browsing. There are also cases in which search engine results were grouped into pre-existing subject categories for browsing. For example, Pratt (1997) who experimented with organizing search results into MeSH categories.
Other mixed approaches are also possible, such as the one applied in the Scorpion project (see Section 2.3.1.2).
The emergence of this approach demonstrates the potentials for utilizing ideas and methods from another community's approach.

3. Discussion
3.1 Features of automated classification approaches
Several problems with automated classification in general have been identified in the literature. As Svenonius (2000, pp. 46-9) claims, automating subject determination belongs to logical positivism – a subject is considered to be a string occurring above a certain frequency, is not a stop word and is in a given location, such as a title.

In clustering algorithms, inferences are made such as "if document A is on subject X, then if document B is sufficiently similar to document A (above a certain threshold), then document B is on that subject." It is assumed that concepts have names, which is common in science, but is not always the case in humanities and social sciences. Automated classification in certain domains has been entirely unexplored, due to lack of suitable data collections or good-quality controlled vocabularies. Another critique given is the lack of theoretical justifications for vector manipulations, such as the cosine measure that is used to obtain vector similarities (Salton, 1991, p. 975).
In regards to similarities and differences between the approaches, document pre-processing (e.g. selection of terms) is common to all the approaches. Various web page characteristics have also been explored by all the three communities, although mostly within the text categorization approach. Major differences between the three approaches are in applied algorithms, employment or not of the vector-space model and of controlled vocabularies, especially as to how well-suited they are for subject browsing (cf. 3.3 Application for subject browsing). Since, there are similarities between approaches, the hypothesis is that idea exchange and co-operation between the three communities would be beneficial. The hypothesis does seem to be supported by the emergence of the mixed approach. They could all benefit from at least looking into each other's approaches to document pre-processing and indexing, and exchanging ideas about properties of web pages and how they could be used. However, there seems to be little co-operation or idea exchange among them. This is also supported by the fact that, to the author's knowledge, no review paper on automated classification attempted to discuss more than one community's approach. A recent bibliometric study (Golub and Larsen, 2005) shows that the three communities are quite clearly mutually independent when looking at citation patterns; and that document clustering and text categorization are closer to each other, while the document classification community is almost entirely isolated. Further research is needed to determine why direct and indirect links are lacking between the document classification and the other two communities, in spite of emergence of the mixed approach.

3.2 Evaluation

The problem of deriving the correct interpretation of a document's subject matter has been much discussed in the library science literature (while much less so in machine learning and information retrieval communities). 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 (Olson and Boll, 2001, pp. 99-101). There are two main factors that seem to affect it:
(1) higher specificity and higher exhaustivity both lead to lower consistency (indexers choose the same first term for the major subject of the document, but the consistency will decrease as they choose more terms); and
(2) the bigger the vocabulary, or, the more choices the indexers have, the less likely they will choose the same terms (Olson and Boll, 2001, pp. 99-101).
The document collection's purpose is another important factor in deciding which classes or terms are to be chosen or made more prominent.
Having the above in mind, performance measures need to be questioned and evaluation has to be dealt with in the broader contexts of users and their tasks. Subject experts or intended end-users have been mostly excluded from evaluation in text categorization and document clustering approaches, while the document classification approach tends to involve them to a larger degree, corresponding to the tradition of evaluating other library services.
Owing to poor evaluation, it is difficult to estimate to what degree the automated classification tools of today are really applicable in operative information systems and for which tasks.

3.3 Application for subject browsing

Research in text categorization seems to be mainly in improving categorization performance, and experiments are conducted under controlled conditions. Research in which web pages have been categorized into hierarchical structures for browsing generally does not involve well-developed classification schemes, but home-grown structures such as directories of search engines that are not structured and maintained well enough.

In document clustering, clusters' labels and relationships between the clusters are automatically produced. Labelling of the clusters is a major research problem, with relationships between the categories, such as those of equivalence, related-term and hierarchical relationships, being even more diffi to automatically derive (Svenonius, 2000, p. 168). "Automatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand" (Chen and Dumais, 2000). Also, clusters change as new documents are added to the collection. Unstable category names in web services and digital libraries, for example, are not user-friendly. Koch and Zettergren (1999) suggest that document clustering is better suited for organizing web search engine results.

Document classification approach employs well-developed classification schemes, which are suitable for subject browsing. However, future research should include improving controlled vocabularies for browsing in the electronic environment, as well as making them more suitable for automated classification.

References

20 Newsgroups DataSet (1998), The 4 Universities Data Set, available at: www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html (accessed 22 December 2004).

DDC (2005), "About DDC: research: a vital part of ongoing development", Dewey Services, available at: www.oclc.org/dewey/about/research/ (accessed 8 August 2005).

Ardo¨, A. et al., (1994), "Improving resource discovery and retrieval on the internet: the Nordic WAIS/world wide web project summary report", NORDINFO Nytt, Vol. 17 No. 4, pp. 13-28.

Attardi, G., Gull`ı, A. and Sebastiani, F. (1999), "Automatic web page categorization by link and context analysis", in Hutchison, C. and Lanzarone, G. (Eds), Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence, pp. 105-19.

Bekkerman, R. et al., (2003), "Distributional word clusters vs words for text categorization", Journal of Machine Learning Research, Vol. 3, pp. 1183-208.

BINDEX (2001), "HLT Project Factsheet: BINDEX", HLTCentral, available at: www.hltcentral.org/projects/print.php?acronym ¼ BINDEX (accessed 22 December 2004).

Blum, A. and Mitchell, T. (1998), "Combining labeled and unlabeled data with co-training", COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers, San Mateo, CA.

Cai, L. and Hofmann, T. (2003), "Text categorization by boosting automatically extracted concepts", in Callan, J. et al. (Eds), Proceedings of SIGIR-03, 26th ACM International Conference on Research and Development in Information Retrieval, pp. 182-9.

CERES (2003), "CERES thesaurus effort", CERES The California Environmental Resources Evaluation System, available at: http://ceres.ca.gov/thesaurus/ (accessed 22 December 2004).

Chakrabarti, S. et al. (1998a), "Automatic resource compilation by analyzing hyperlink structure and associated text", Proceedings of the Seventh International Conference on World Wide Web 7, Brisbane, Australia, pp. 65-74.

Chakrabarti, S., Dom, B. and Indyk, P. (1998b), "Scalable feature selection, classification and signature

generation for organizing large text databases into hierarchical topic taxonomies", Journal of Very Large Data Bases, Vol. 7 No. 3, pp. 163-78.
Chan, L.M. (1994), Cataloging and Classification: An Introduction, 2nd ed., McGraw-Hill, New York, NY.
Chen, H. and Dumais, S.T. (2000), "Bringing order to the web: automatically categorizing search results", Proceedings of CHI-00, ACM International Conference on Human Factors in Computing Systems, Den Haag, pp. 145-52.
Chen, M., LaPaugh, A. and Singh, J.P. (2002), "Categorizing information objects from user access patterns", Proceedings of the Eleventh International Conference on Information and Knowledge Management, 4-9 November, pp. 365-72.
Clusty (2004), "Clusty the clustering engine", Vivsimo, available at: www.clusty.com (accessed 22 December 2004).
Cutting, D. et al. (1992), "Scatter/gather: a cluster-based approach to browsing large document collections", Proceedings of the 15th Annual International ACM/SIGIR Conference, Copenhagen, pp. 318-29.
DESIRE (2000), "DESIRE: development of a European service for information on research and education", DESIRE, available at: www.desire.org/ (accessed 22 December 2004).
DESIRE Project (1999), Lunds Universitets Bibliotek, available at: www.lub.lu.se/desire (accessed 22 December 2004).
Dittenbach, M., Berger, H. and Merkl, D. (2004), "Improving domain ontologies by mining semantics from text", Proceedings of the first Asian-Pacific Conference on Conceptual Modeling, Dunedin, New Zealand,Vol. 31, pp. 91-100.
Dumais, S.T. and Chen, H. (2000), "Hierarchical classification of web content", Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 24-28 July, Athens, Greece, pp. 256-63.
Dumais, S.T., Lewis, D.D. and Sebastiani, F. (2002), "Report on the workshop on operational text classification systems (OTC-02)", ACM SIGIR Forum, Vol. 35 No. 2, pp. 8-11.
EELS (2003), "'All' Engineering resources on the internet: a companion service to EELS", EELS, Engineering E-Library, Sweden, available at: http://eels.lub.lu.se/ae/ (accessed 22 December 2004).
Engine-e (2004), Lund University Libraries, available at: http://engine-e.lub.lu.se/ (accessed 22 December).
Engineering Electronic Library (2003), Lund University Libraries, available at: http://eels.lub.lu. se/ (accessed 22 December 2004).
FAST (2003), "FAST as a knowledge base for automated classification", OCLC projects, available at: www.oclc.org/research/projects/fastac/ (accessed 7 August 2005).
FAST (2004), "FAST: faceted application of subject terminology", OCLC projects, available at: www.oclc.org/research/projects/fast/ (accessed 22 December 2004).
Fasulo, D. (1999), "An analysis of recent work on clustering algorithms: technical report", University of Washington, available at: http://citeseer.nj.nec.com/fasulo99analysi.html (accessed 22 December 2004).
Fisher, M. and Everson, R. (2003), "When are links useful? Experiments in text classification", Proceedings of ECIR-03, 25th European Conference on Information Retrieval, Pisa, IT, pp. 41-56.
Frank, E. and Paynter, G.W. (2004), "Predicting library of congress classifications from library of congress subject headings", Journal of the American Society for Information Science and Technology, Vol. 55 No. 3, pp. 214-27.
Fu¨ rnkranz, J. (1999), "Exploiting structural information for text classification on the WWW",

Proceedings of IDA-99, 3rd Symposium on Intelligent Data Analysis, pp. 487-97.
Fu¨ rnkranz, J. (2002), “Hyperlink ensembles: a case study in hypertext classification”, Information Fusion, Vol. 3 No. 4, pp. 299-312.
Garfield, E., Malin, M.V. and Small, H. (1975), “A system for automatic classification of scientific literature”, Journal of the Indian Institute of Science, Vol. 57 No. 2, pp. 61-74, (Reprinted in: Essays of an Information Scientist, Vol. 2, pp. 356-65).
GERHARD (1998), “GERHARD: German harvest automated retrieval and directory”, GERHARD, available at: www.gerhard.de/ (accessed 22 December 2004).
GERHARD (1999), “GERHARD – navigating the web with the universal decimal classification system”, GERHARD, available at: www.gerhard.de/info/dokumente/vortraege/ecdl99/ html/index.htm (accessed 22 December 2004).
Ghani, R., Slattery, S. and Yang, Y. (2001), “Hypertext categorization using hyperlink patterns and metadata”, Proceedings of ICML-01, 18th International Conference on Machine Learning, pp. 178-85.
Glover, E.J. et al. (2002), “Using web structure for classifying and describing web pages”, Proceedings of the Eleventh International Conference on World Wide Web Honolulu, Hawaii, USA, pp. 562-9.
Glover, E.J. et al. (2003), “Inferring hierarchical descriptions”, Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM 2002, November 4-9, pp. 507-14.
Godby, J. and Reighart, R. (1998), “The WordSmith indexing system”, OCLC Digital Archive, available at: http://digitalarchive.oclc.org/da/ViewObject.jsp?fi ¼ 0000003487: 000000090408&reqid ¼ 33836 (accessed 22 December 2004).
Golub, K. and Larsen, B. (2005), “Different approaches to automated classification: is there an exchange of ideas?”, in Ingwersen, P. and Larsen, B. (Eds), Proceedings of ISSI 2005 – the 10th International Conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, 24-28 July,Vol. 1, Karolinska University Press, Stockholm, pp. 270-4.
Goren-Bar, D. et al. (2000), “Supervised learning for automatic classification of documents using self-organizing maps”, Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, ZU˝ rich, Switzerland, Vol. 11-12, p. 2000.
Go¨vert, N., Lalmas, M. and Fuhr, N. (1999), “A probabilistic description-oriented approach for categorising web documents”, Proceedings of the Eighth International Conference on Information and Knowledge Management, pp. 475-82.
Hartigan, J.A. (1996), “Introduction”, in Hubert, L. and De Soete, G. (Eds), Clustering and Classification Arabie, World Scientific, Singapore.
Hatzivassiloglou, V., Gravano, L. and Maganti, A. (2000), “An investigation of linguistic features and clustering algorithms for topical document clustering”, Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 224-31.
Haveliwala, T.H., Gionis, A. and Indyk, P. (2000), “Scalable techniques for clustering the web”, Third International Workshop on the Web and Databases, May, pp. 129-34.
Hersh, W.R. (1994), “OHSUMED: an interactive retrieval evaluation and new large test collection for research”, Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 192-201.
Heuser, U., Babanine, A. and Rosenstiel, W. (1998), “HTML documents classification using (non-linear) principal component analysis and self-organizing maps”, Proceedings of the Fourth International Conference on Neural Networks and their Applications (Neurap’98), 11-13 March 1998, Marseilles, France, pp. 291-5.

INitiative for the Evaluation of XML Retrieval (2004), DELOS Network of Excellence for Digital Libraries, available at: http://inex.is.informatik.uni-duisburg.de/ (accessed 22 December 2004).
Jain, A.K., Murty, M.N. and Flynn, P.J. (1999), "Data clustering: a review", ACM Computing Surveys, Vol. 31 No. 3, pp. 264-323.
Jenkins, C. et al., (1998), "Automatic classification of web resources using Java and Dewey decimal classification", Computer Networks & ISDN Systems, Vol. 30, pp. 646-8.
Kim, H.R. and Chan, P.K. (2003), "Learning implicit user interest hierarchy for context in personalization", Proceedings of the International Conference on Intelligent User Interfaces, pp. 101-8.
Koch, T. (1994), "Experiments with automatic classification of WAIS databases and indexing of WWW", Internet World & Document Delivery World International 94, London, May, pp. 112-5.
Koch, T. and Ardo¨, A. (2000), "Automatic classification", DESIRE II D3.6a, Overview of Results, available at: www.lub.lu.se/desire/DESIRE36a-overview.html (accessed 22 December 2004).
Koch, T. and Day, M. (1997), "The role of classification schemes in internet resource description and discovery", EU Project DESIRE, Deliverable D3.2.3, available at: www.lub.lu.se/desire/radar/reports/D3.2.3/ (accessed 22 December 2004).
Koch, T. and Zettergren, A-S. (1999), "Provide browsing in subject gateways using classification schemes", EU Project DESIRE II, available at: www.lub.lu.se/desire/handbook/class.html (accessed 22 December 2004).
Kohonen, T. (2001), Self-Organizing Maps, 3rd ed., Springer-Verlag, Berlin.
Koller, D. and Sahami, M. (1997), "Hierarchically classifying documents using very few words", Proceedings of ICML-97, 14th International Conference on Machine Learning, pp. 170-8.
Labrou, Y. and Finin, T. (1999), "Yahoo! As an ontology: using Yahoo! Categories to describe documents", Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management, pp. 180-7.
Larson, R.R. (1992), "Experiments in automatic library of congress classification", Journal of the American Society for Information Science, Vol. 43 No. 2, pp. 130-48.
Li, Y.H. and Jain, A.K. (1998), "Classification of text documents", The Computer Journal, Vol. 41 No. 8, pp. 537-46.
Liere, R. and Tadepalli, P. (1998), "Active learning with committees: preliminary results in comparing winnow and perception in text categorization", Proceedings of CONALD-98, 1st Conference on Automated Learning and Discovery.
Lindholm, J., Scho¨nthal, T. and Jansson, K. (2003), "Experiences of harvesting web resources in engineering using automatic classification", Ariadne, No. 37, available at: www.ariadne.ac.uk/issue37/lindholm/
Liu, X. et al. (2002), "Document clustering with cluster refinement and model selection capabilities", Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 191-8.
McCallum, A. et al. (1998), "Improving text classification by shrinkage in a hierarchy of classes", paper presented at ICML-98, 15th International Conference on Machine Learning, pp. 359-67.
McCallum, A. et al. (1999), "Building domain-specific search engines with machine learning techniques", paper presented at AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace.
McCallum, A. et al., (2000), "Automating the construction of internet portals with machine learning", Information Retrieval Journal, Vol. 3, pp. 127-63.

Mandhani, B., Joshi, S. and Kummamuru, K. (2003), "A matrix density based algorithm to hierarchically co-cluster documents and words", Proceedings of the Twelfth International Conference on World Wide Web, Budapest, Hungary, pp. 511-8.
Manning, C. and Schu¨ tze, H. (1999), Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA.
Merchkour, M., Harper, D.J. and Muresan, G. (1998), "The WebCluster project: using clustering for mediating access to the world wide web", Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 357-8.
MetaCrawler Web Search (2005), available at: http://metacrawler.com (accessed 5 August 2005).
Mitchell, T. (1997), Machine Learning, McGraw-Hill, New York, NY.
Mladenic, D. (1998), "Turning Yahoo into an automatic web-page classifier", Proceedings of the 13th European Conference on Artificial Intelligence ECAI'98, pp. 473-4.
Mladenic, D. and Grobelnik, M. (2003), "Feature selection on hierarchy of web documents", Decision Support Systems, Vol. 35 No. 1, pp. 45-87.
Mo¨ller, G. et al. (1999), "Automatic classification of the WWW using the universal decimal classification", in McKenna, B. (Ed.), Proceedings of the 23rd International Online Information Meeting, London, 7-9 December, pp. 231-8.
Nordic WAIS/World Wide Web Project (1995), Lund University Libraries, available at: www.lub. lu.se/W4/ (accessed 22 December 2004).
Nu¨ bel, R. et al. (2002), "Bilingual indexing for information retrieval with AUTINDEX", LREC Proceedings, Las Palmas.
Olson, H.A. and Boll, J.J. (2001), Subject Analysis in Online Catalogs, 2nd ed., Libraries Unlimited, Englewood, CO.
Palmer, C.R. et al. (2001), "Demonstration of hierarchical document clustering of digital library retrieval results", Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, Virginia, p. 451.
Pierre, J.M. (2001), "On the automated classification of web sites", Linko¨ ping Electronic Articles in Computer and Information Science, Vol. 6 No. 001.
Poincot, P., Lesteven, P.S. and Murtagh, F. (1998), "A spatial user interface to the astronomical literature", Astronomy & Astrophysics, 2 May, pp. 183-91.
Pratt, W. (1997), "Dynamic organization of search results using the UMLS", American Medical Informatics Association Fall Symposium, pp. 480-4.
Rasmussen, E. (1992), "Clustering algoritms", in Frakes, W.B. and Baeza-Yates, R. (Eds), Information Retrieval: Data Structures and Algorithms, Prentice-Hall, Engelwood Cliffs, NJ.
Rauber, A. and Merkl, D. (1999), "SOMLib: a digital library system based on neural networks", Proceedings of the Fourth ACM Conference on Digital Libraries, Berkeley, California, United States, pp. 240-1.
Reuters-21578 (2004), available at: www.daviddlewis.com/resources/testcollections/ reuters21578/ (accessed 3 August 2005).
Rocchio, J.J. (1971), "Relevance feedback in information retrieval", in Salton, G. (Ed.), The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, Englewood Cliffs, NJ, pp. 313-23.
Ruiz, M.E. and Srinivasan, P. (1999), "Hierarchical neural networks for text categorization", Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 281-2.

Sahami, M., Yusufali, M. and Baldonado, M.Q. (1998), "SONIA: a service for organizing networked information autonomously", paper presented at 3rd ACM Conference on digital libraries, Pittsburgh, pp. 200-9.
Salton, G. (1991), "Developments in automatic text retrieval", Science, Vol. 253, pp. 974-9.
Schu¨ tze, H., Hull, D.A. and Pedersen, J.O. (1995), "A comparison of classifiers and document representations for the routing problem", Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, pp. 229-37.
Schwartz, C. (2001), Sorting Out the Web: Approaches to Subject Access, Ablex, Westport, CT.
Schweighofer, E., Rauber, A. and Dittenbach, M. (2001), "Automatic text representation, classification and labeling in European law", ICAIL 2001, pp. 78-87.
Scorpion (2004), OCLC software, available at: www.oclc.org/research/software/scorpion/default. htm (accessed 22 December).
Sebastiani, F. (2002), "Machine learning in automated text categorization", ACM Computing Surveys, Vol. 34 No. 1, pp. 1-47.
Slattery, S. and Craven, M. (2000), "Discovering test set regularities in relational domains", Proceedings of ICML-00, 17th International Conference on Machine Learning, pp. 895-902.
Slonim, N., Friedman, N. and Tishby, N. (2003), "Unsupervised document classification using sequential information maximization", Proceedings of SIGIR'02, 25th ACM International Conference on Research and Development of Information Retireval, Tampere, Finland, 2002.
Soergel, D. et al., (2004), "Reengineering thesauri for new applications: the AGROVOC example", Journal of Digital Information, Vol. 4 No. 4, Article No. 257, available at: http://jodi.ecs. soton.ac.uk/Articles/v04/i04/Soergel/
Steinbach, M., Karypis, G. and Kumar, V. (2000), "A comparison of document clustering techniques", KDD Workshop on Text Mining, Boston, MA, 20-23 August.
Su, Z. et al. (2001), "Correlation-based document clustering using web logs", Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34), 3-6 January,Vol. 5, p. 5022.
Subramanian, S. and Shafer, K.E. (1998), "Clustering", OCLC Publications, available at: http:// digitalarchive.oclc.org/da/ViewObject.jsp?objid ¼ 0000003409 (accessed 22 December 2004).
Sun, A., Lim, E-P. and Ng, W-K. (2001), "Hierarchical text classification and evaluation", ICDM 2001, IEEE International Conference on Data Mining.
Svenonius, E. (2000), The Intellectual Foundations of Information Organization, MIT Press, Cambridge, MA.
Thunderstone (2005), Thunderstone's Web Site Catalog, available at: http://search.thunderstone. com/texis/websearch (accessed 4 August 2005).
Tombros, A. and van Rijsbergen, C.J. (2001), "Query-sensitive similarity measures for the calculation of interdocument relationships", Proceedings of the Tenth International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, pp. 17-24.
Toth, E. (2002), "Innovative solutions in automatic classification: a brief summary", Libri, Vol. 25 No. 1, pp. 48-53.
TREC (2004), "TREC: Text REtrieval Conference", National Institute of Standards and Technology, available at: http://trec.nist.gov/ (accessed 22 December 2004).
Vizine-Goetz, D. (1996), "Using library classification schemes for internet resources", OCLC Internet Cataloging Project Colloquium, available at: http://webdoc.sub.gwdg.de/ebook/ aw/oclc/man/colloq/v-g.htm, (accessed 4 April 2006).
Wacholder, N., Evans, D.K. and Klavans, J.L. (2001), "Automatic identification and organization of

index terms for interactive browsing", Proceedings of the ACM-IEEE Joint Conference on Digital Libraries, Roanoke, Virginia, June, pp. 128-34.
Wallis, J. and Burden, P. (1995), "Towards a classification-based approach to resource discovery on the web", University of Wolverhampton, Wolverhampton, available at: www.scit.wlv. ac.uk/wwlib/position.html (accessed 22 December 2004).
Wang, Y. and Kitsuregawa, M. (2002), "Evaluating contents-link coupled web page clustering for web search results", Proceedings of the Eleventh International Conference on Information and Knowledge Management, McLean, Virginia, USA, pp. 499-506.
WebKB (2001), CMU World Wide Knowledge Base, available at: www-2.cs.cmu.edu/ , webkb/ (accessed 22 December 2004).
Weiss, R. et al. (1996), "HyPursuit: a hierarchical network search engine that exploits content-link hypertext clustering", Proceedings of the Seventh ACM Conference on Hypertext, Washington, DC, March, pp. 180-93.
Willet, P. (1988), "Recent trends in hierarchic document clustering: a critical review", Information Processing and Management, Vol. 24 No. 5, pp. 577-97.
Yahoo! (2005), Yahoo! Directory, available at: http://dir.yahoo.com/ (accessed 8 August 2005). Yang, Y. (1999), "An evaluation of statistical approaches to text categorization", Journal of Information Retrieval, Vol. 1 Nos 1/2, pp. 67-88.
Yang, C., Chen, H. and Hong, K. (2003), "Visualization of large category map for internet browsing", Decision Support Systems (DSS), Vol. 35 No. 1, pp. 89-102.
Yang, Y., Slattery, S. and Ghani, R. (2002), "A study of approaches to hypertext categorization", Journal of Intelligent Information Systems, Vol. 8 Nos 2/3, pp. 219-41.
Zamir, O. and Etzioni, O. (1998), "Web document clustering: a feasibility demonstration", ACM SIGIR'98, Australia, pp. 46-54.
Zamir, O. et al. (1997), "Fast and intuitive clustering of web documents", Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, pp. 287-90.
Zhao, Y. and Karypis, G. (2002), "Evaluation of hierarchical clustering algorithms for document dataset", Proceedings of the Eleventh International Conference on Information and Knowledge Management, McLean, Virginia, pp. 515-24.

About the author

Koraljka Golub is a PhD student at the Department of Information Technology, Lund University, Sweden. She has a Master's degree in Library and Information Science from the University of Zagreb, Croatia. Within the field of digital libraries, she is interested in automated subject classification and user information behaviour related to subject browsing. Koraljka Golub can be contacted at: koraljka.golub@it.lth.se

Many thanks to Traugott Koch, Anders Ardo¨ Tatjana Aparac Jelusˇic´ Johan Eklund, Ingo Frommholz, Repke de Vries and the Journal of Documentation reviewers for providing valuable feedback on earlier versions of the paper.