Automatic Subject Indexing of Text†

Koraljka Golub

Linnaeus University, School of Cultural Sciences,
Department of Library and Information Science,
Faculty of Arts and Humanities, 351 95 Växjö, Sweden,
<koraljka.golub@lnu.se>

Koraljka Golub is an associate professor in library and information science at Linnaeus University, Sweden. Her research interests focus on knowledge organization, primarily in the context of information retrieval. Research projects she has worked on have explored the potential of social tagging when enhanced by suggestions from controlled vocabularies, automatic subject indexing, and evaluation of subject indexing in the context of retrieval. She would like to examine to what degree automatic full-text indexing, end-user tagging, author tagging, professional subject indexing, and automatic assigned indexing, or any combination thereof, contribute to successful retrieval.

Golub, Koraljka. 2019. "Automatic Subject Indexing of Text." *Knowledge Organization* 46(2): 104-121. 126 references. DOI:10.5771/0943-7444-2019-2-104.

Abstract: Automatic subject indexing addresses problems of scale and sustainability and can be at the same time used to enrich existing metadata records, establish more connections across and between resources from various metadata and resource collections, and enhance consistency of the metadata. In this work, automatic subject indexing focuses on assigning index terms or classes from established knowledge organization systems (KOSs) for subject indexing like thesauri, subject headings systems and classification systems. The following major approaches are discussed, in terms of their similarities and differences, advantages and disadvantages for automatic assigned indexing from KOSs: "text categorization," "document clustering," and "document classification." Text categorization is perhaps the most widespread, machine-learning approach with what seems generally good reported performance. Document clustering automatically both creates groups of related documents and extracts names of subjects depicting the group at hand. Document classification re-uses the intellectual effort invested into creating a KOS for subject indexing and even simple string-matching algorithms have been reported to achieve good results, because one concept can be described using a number of different terms, including equivalent, related, narrower and broader terms. Finally, applicability of automatic subject indexing to operative information systems and challenges of evaluation are outlined, suggesting the need for more research.

Received: 27 September 2018; Revised: 31 October 2018; Accepted: 7 December 2018

Keywords: indexing, subject, terms, document, documents, automatic, classification

† Many thanks to Birger Hjørland and two anonymous reviewers who kindly provided detailed feedback that helped improve this article.

1.0 Introduction

Increasingly, different types of information resources are being made available online. Current search engines yield good results for specific search tasks but are unsuited to the conceptual or subject-based searches requiring high precision and recall, common in academic research or serious public inquiry (for a discussion on (dis)advantages of automatic full-text indexing, see Keyser 2012, chapter 2). Differences in terminology between various communities and even individuals lead to the fact that literal string search in many cases cannot deliver effective search. This is exacerbated in cross-system and cross-lingual search and retrieval where integrated subject access is probably the hardest challenge to address. Subject index terms taken from knowledge organization systems (KOSs) such as thesauri, subject headings systems and classification systems provide numerous benefits compared to the free-text indexing of commercial search engines: consistency through uniformity in term format and assignment of terms, provision of semantic relationships among terms, and support for browsing through consistent and clear hierarchies (see Mazzocchi 2018).

However, such subject index terms require substantial resources to produce. Because of the ever-increasing number of documents, there is a risk that recognized objectives of bibliographic systems, such as finding all documents on a given subject, would get left behind. As an example, a recent exploratory study of Swedish library catalogs indicates that subject access is not addressed systematically, that in new digital collections KOSs are applied to a very limited degree, and in integrated library and commercial databases the mappings between the different KOSs do not exist, therefore preventing quality search across them (Golub 2016). Automatic means could be a solution to preserve recognized objectives of bibliographic systems

(Svenonius 2000, 30). Apart from addressing problems of scale and sustainability, automatic subject indexing can be used to enrich existing bibliographic records, establish more connections across and between resources, and enhance consistency of bibliographic data (Golub et al. 2016). Further, automatic indexing is used today in a wide variety of applications such as topical harvesting, personalized routing of news articles, ranking of search engine results, sentiment analysis (see, e.g., Hu and Li 2011) and many others (Sebastiani 2002).

Research on automatic subject indexing began with the availability of electronic text in the 1950s (Luhn 1957; Baxendale 1958; Maron 1961) and continues to be a challenging topic, for the reasons and purposes outlined above. For a historical overview of automatic indexing, see Stevens (1965) and Sparck Jones (1974) covering the early period of automatic indexing and Lancaster (2003, 289-292) for the later one. A related term is machine-aided indexing (MAI) or computer-assisted indexing (CAI) where it is the human indexer who decides, based on a suggestion provided by the computer (see, for example, Medical Text Indexer (U.S. National Library of Medicine, 2016)). A similar approach is applied by Martinez-Alvarez, Yahyaei, and Roelleke (2012) who propose a semi-automatic approach in which only those predictions likely to be correct are processed automatically, while more complex decisions are left to human experts to decide.

There are different approaches to automatic indexing, based on the purpose of application but also coming from different research fields and traditions. The terminology is, therefore, varied. Further, research of automatic indexing tools in operating information environments is usually conducted in laboratory conditions, excluding the complexities of real-life systems and situations. The remainder of this entry reflects upon these issues and is structured as follows: the next section (2) discusses major terms and provides definition of automatic subject indexing as used for the purposes of this work. Section 3 discusses approaches to automatic subject indexing as to their major similarities and differences. Section 4 contains a discussion on how good the addressed automatic solutions are today, and Section 5 contains concluding remarks.

2.0 Definition and terminology

According to the current ISO indexing standard (ISO 5963:1985, confirmed in 2008, International Organization for Standardization 1985), subject indexing performed by the information professional is defined as a process involving three steps: 1) determining the subject content of a document; 2) a conceptual analysis to decide which aspects of the content should be represented; and, 3) translation of those concepts or aspects into a controlled vocabulary (CV). Automatic subject indexing is, then, a machine-based subject indexing where human intellectual processes of the above three steps are replaced by, for example, statistical and computational linguistics techniques, which will be discussed in further detail below.

The terminology related to automatic subject indexing is inconsistently used in the literature. This is probably because this research topic has been addressed by different research fields and disciplines, grounded in various epistemological traditions. In order to clarify the differences, major terms used are briefly discussed and defined below.

In information science, the terminology of subject indexing involves several important concepts. Subject index terms may be derived either from the document itself, which is known as derived indexing (e.g., keywords taken from title), or from indexing languages that are formalized and specifically designed for describing the subject content of documents, which is known as assigned indexing or classification. In assigned indexing, index terms are taken from alphabetical indexing languages (using natural language terms with terminology control such as thesauri and subject headings); in classification, classes are taken from classification systems (using symbols, operating with concepts). The main purpose of assigned indexing using alphabetical indexing languages is to allow retrieval of a document from many different perspectives; typically, three to twenty elemental or moderately pre-combined subject terms are assigned. The main purpose of classification, assigning classes from classification schemes, is to group similar documents together to allow browsing (of library shelves in the traditional environment and directory-style browsing in the online environment); a few, typically one, highly pre-combined subject class(es) are assigned. (See also Lancaster (2003, 20-21) concerning the similarities between indexing and classification).

In computer science, the distinction between different types of indexing languages is rarely made. While a common distinction made is the one between formal ontologies, light ontologies (with concepts connected using general associative relations rather than strict formal ones typical of the former) and taxonomies, at times the term ontology is used to refer to several different knowledge organization systems. For example, Mladenić and Grobelnik use the term to refer to hierarchical web directories of search engines and related services as well as subject headings systems (2005, 279):

> Most of the existing ontologies were developed with considerable human efforts. Examples are Yahoo! and DMOZ topic ontologies containing Web pages or MESH ontology of medical terms connected to Medline collection of medical papers.

Also, derived indexing may be variously termed, for example keyword assignment, keyword extraction, or noun phrase extraction (referring to noun phrases specifically).

In related literature, other terms for automatic subject indexing are used. Subject metadata generation is one general example. Terms text categorization and text classification are common in the machine learning community. Automatic classification is another example of a term used to denote automatic assignment of a class or a category from a pre-existing classification system or taxonomy. However, this phrase may also be used to refer to document clustering, in which groups of similar documents are automatically discovered and named.

Here the term automatic subject indexing is used as the primary term. It denotes non-intellectual, machine-based processes of subject indexing as defined by the information science community: derived and assigned indexing using both alphabetical and classification indexing systems for the purposes of improved information retrieval. The rationale for combining them into one entry is the fact that the underlying machine-based principles are rather similar, especially when it comes to application to textual documents. However, the major focus in this entry is on assigned indexing because of the added value provided by indexing systems for information searching as perceived in information science, such as increased precision and recall ensuing from natural language control of, e.g., homonymy, synonymy, word form, and advantages for hierarchical browsing, e.g., when the end-user does not know which search term to use because of unfamiliarity with the topic or when not looking for a specific item. Further, term subject indexing assumes applying both alphabetical and classification indexing systems, because similar principles apply when it comes to automatic processes; although, it is also common to refer to the process of using the former subject indexing and the latter subject classification. Finally, while the word automated more directly implies that the process is machine-based, the word automatic is more commonly used in related literature and has, therefore, become the term of choice here, too.

Further, terminology to distinguish between different approaches to automatic subject indexing is even less consistent (see also Smiraglia and Cai 2017). For example, Hartigan (1996, 2) writes: "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 Schütze 1999, 575). In this entry, terms text categorization and document clustering are chosen, because they tend to be the prevalent terms in the literature of the corresponding communities. Term document classification is used in order to consistently distinguish between the three approaches. These approaches are described and discussed in the following section.

3.0 Approaches to automatic subject indexing

This section (3.1) first describes the underlying methodology common in different specific approaches. Section 3.2 provides a brief overview of addressing various document types. Section 3.3 discusses the major approaches, text categorization, document clustering and document classification.

3.1 Basic approach

Generally speaking, automatic subject indexing typically follows a course of several major steps. The first one is a preparation step in which documents to be indexed are each processed in order to create suitable representations for computer manipulation in what follows. This process is comparable to preparation of documents for information retrieval.

3.1.1 Pre-processing

A list of words appearing in the document is created based on tokenization, the process of automatically recognizing words. Also, all punctuation is taken away. Further, words that tend to carry less meaning are taken out, such as conjunctions, determiners, prepositions, and pronouns, all of which are known as stop-words. This resulting representation of documents is known as a bag-of-words model. A more advanced representation is the n-gram model of words which is used, for example, when noun phrases need to be extracted in derived indexing or when string matching is conducted against terms comprising more than just one word (see below section 3.3.3. Document classification). Word n-grams may be unigrams (individual words), bigrams (any two adjacent words), trigrams (any three adjacent words), etc. Further, more advanced natural language processing techniques may be performed; in stemming, each word is reduced to its stem, which means removal of its affixes—for example, illegally may be reduced to its stem legal whereby its prefix il- and its suffix -ly are removed. The rationale behind this is that words with the same stem bear the same meaning. In addition, part-of-speech taggers and syntactic parsers can also be applied. For an overview of text processing, see Manning and Schütze (1999) and Weisser (2015).

3.1.2 Term weighting

The following major step is determining the importance of each term for describing the aboutness of the document at hand. The term can be either an individual word or a compound phrase, depending on the given task. For each term, a weight expressed as a number is calculated and assigned. Here different statistical and other heuristic rules can be applied. An example of statistical rules, words appearing very many times both in the document at hand and in all other documents in the collection, are probably not particularly indicative of the subject matter of the document and vice versa. This is known as term frequency-inverse document frequency weight (tf-idf, Salton and McGill 1983, 63, 205): it combines 1) term frequency (Luhn 1957), where weight of the term at hand is considered to be proportional to the number of times it appears in the document, with 2) inverse document frequency (Sparck Jones 1972), where weight of the term is an inverse fraction of the documents that contain the word. An overview of term weighting measures can be found in Roelleke (2013).

Features such as the location of the term, or the font size or font type, may also be included in determining the importance of a term. In web pages, for example, words that appear in titles, headings or metadata may be considered more indicative of the topicality than those written in normal font size elsewhere. A known example is Google that owes much of its success to the PageRank algorithm (Page et al. 1998) that ranks higher those web pages which have more external web pages linking to them. Gil-Leiva (2017) pointed out that generally there is less use of location heuristics rules than of statistical rules (outlined in the previous paragraph) and conducted an experiment comparing the two sets of rules, which showed that best results are achieved with location heuristics rules. A number of other principles have also been investigated. A co-occurrence, or a citation-based one applies the idea that if publication A cites publication B, A may include text that indicates what B is about (Bradshaw and Hammond 1999). Chung, Miksa, and Hastings (2010) compared how sources human indexers normally resort to in order to determine the subject of the document at hand, such as conclusion, abstract, introduction, title, full text, cited works, and keywords of scientific articles, contribute to automatic indexing performance. Using the SVM implementation in Weka (Witten and Frank 2000), they gained results that indicated keywords outperformed full-text, while cited works, source title (title of journal or conference), and title were all as effective as the full text.

Rules can be of different types. Driscoll et al. (1991) matched the document text against over 3,000 phrases and a set of deletion and insertion rules. These rules were used to transform the list of terms from the document to the list of index phrases; for example, if "time," "over," and "target" appeared within a certain number of words from each other, an index phrase "air warfare" would be generated. Fuhr and Knorz (1984) created about 150,000 rules for matching physics documents to KOS terms. Jones and Bell (1992) extracted index terms based on matching terms from the document against several lists: a stop-word list, a list of terms of interest, a list to aid in the disambiguation of homographs, a list to conflate singular and plural forms, and a list of word endings to allow simple parsing. Ruiz, Aronson, and Hlava (2008) claim that rule-based approaches dominated in the 1970s and 1980s and that machine learning or statistical approaches picked up in the 1990s. Rule-based approaches are based on manually created rules while in machine learning sets of examples are required for training the algorithm to learn concepts. Hlava (2009) describes rule-based indexing as better and states that the majority of rules are simple and can be automatically created, while complex rules are added by editors. On the other hand, in the domain of medical documents, Humphrey et al. (2009) compared a rule-based and statistical approach and showed that the latter outperformed the former. Approaches combining the best of the two worlds may be superior.

3.1.3 Further representations

Based on the two aforementioned major commonly applied processes, each original document is now transformed into a list of (stemmed, parsed) terms and their assigned term weights. There seem to be two possible ways to continue from here: a) vector representation; or, b) string matching.

a. Vector representation is the dominant approach in which the result of the first two steps is now transformed into vectors in a vector space of terms. In this vector space, each term with its weight is represented as one dimension in that space (term space). When features like location are added, each feature becomes a dimension in the vector space called feature space which could then contain the term space. Many terms and features will lead to the challenge of high dimensionality; research has been suggesting dimensionality reduction methods such as: choosing only terms with highest weights, selecting clusters of closest terms instead of terms, taking only parts of documents like summaries or web page snippets. Vector space representation allows for advanced mathematical manipulations beyond what would be possible with just strings of text.
b. Less commonly applied is a string-matching approach between terms from the document and terms describing concepts from an indexing language.

In assigned automatic indexing, a parallel process is taking place to represent target index terms (e.g., classes from a classification system, descriptors from a thesaurus). For example, in subject indexing languages such as thesauri, one concept can be represented by a certain number of synonymous terms, related terms, narrower and broader terms. Or, each concept can be represented by terms extracted from documents that have been manually indexed by the term representing that concept. These representations need to be transformed into vectors when the documents are represented as vectors in order to allow the comparison.

3.1.4 Assignment of index terms

In this final step, either a) vector-based comparisons and calculations (when vectors are used), or b) string matching between terms from the documents and terms representing target index terms are conducted. Usually a list of candidate terms is the first result, from which, then, best candidates are selected also applying various statistical and heuristic rules. One example is to assign the candidate term if it is among the top five and appears in the title of the document, or, more simply, select the top, say, three candidates with the highest weight.

As seen from the four steps above, the dominant basic approach takes into account only terms, rather than concepts or semantic relationship between terms. Taking advantage of relationships in indexing languages like thesauri and ontologies to identify concepts is another possibility (see section 3.3). Also, there are examples that try to approach this problem in other ways; e.g., Huang et al. (2012) who experimented with a measure for identifying concepts by first mapping words from documents to concepts from Wikipedia and WordNet.

Apart from using KOS, other approaches have been suggested. In latent semantic indexing (LSI), perhaps the best-known example, it is assumed that terms that are used in semantically related documents tend to have similar meanings. Based on this assumption, associations between terms that occur in similar documents are calculated, and then concepts for those documents extracted. LSI was first applied in information retrieval for comparing search query terms to documents, at the conceptual rather than literal level (Deerwester et al. 1988; Meng, Lin, and Yu 2011). LSI has been further developed into related approaches, such as probabilistic LSI (pLSI) (Hofmann 2001) and Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). Statistical approaches also try to identify concepts, in particular the ones based on the distributional hypothesis (Harris 1954). According to the hypothesis, words that appear in same contexts tend to have similar meanings. This has been applied in word2vec models (Mikolov et al. 2013; Goldberg and Levy 2014), which apply neural networks to reconstruct contexts of words. Each unique word is assigned a vector and positioned close to vectors representing words, which often appear in similar contexts.

3.2 Document types

While this encyclopedia entry focuses on automatic subject indexing of textual documents, automatic indexing of non-textual or heterogeneous documents shares principles basic to those presented here. For example, multimedia documents like images, sound and video could also be represented by vectors and processed similarly. However, how exactly features of multimedia such as shapes and color distribution need to be selected and processed, is beyond the scope of this entry. For automatic indexing of non-textual resources, the readers may want to refer to Rasmussen Neal (2012).

Another common document type today is data, where automatic categorization is typically applied for prediction purposes (e.g., weather forecast, medical diagnosis, marketing) as opposed to our context of description. Still, many of the principles are similar to ours. For further information, please refer to Kelleher, Mac Namee, and D'Arcy (2015).

When it comes to textual documents, there also are many different sub-types, and while the basic approach described above tends to be applied in most cases, special challenges may arise, as well as special features that could be beneficially explored. For example, web documents have specific characteristics such as hyperlinks and anchors, metadata, and structural information, all of which could serve as complementary features to improve automatic classification. In addition, geographic location, use profiles, citation, and linking, like in PageRank mentioned previously, may be utilized. 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 be misapplied, and titles can be general ("home page," "untitled document") (see, e.g., Gövert, Lalmas, and Fuhr 1999; Golub and Ardö 2005; Klassen and Paturi 2010). Apart from web pages, the following is a non-exhaustive list of textual document examples where research in automatic indexing has been conducted (in no particular order): archival records (e.g., Sousa 2014), doctoral theses (e.g., Hamm and Schenider 2015), clinical medical documents (e.g., Stanfill et al. 2010), e-government (e.g., Svarre and Lykke 2013), business information (Flett and Laurie 2012), online discussions (e.g., Mu et al. 2012), parliamentary resolutions (De Campos and Romero 2008), political texts on the web (Dehghani et al. 2015), grey literature (e.g., Mynarz and Skuta 2011), written documents from businesses like invoices, reminders, and account

statements (e.g., Esser et al. 2012), legal documents for litigation (e.g., Roitblat, Kershaw, and Oot 2010), documents from construction industry such as meeting minutes, claims, and correspondences (e.g., Mahfouz 2012), and documents related to research data such as questionnaires and case studies (El-Haj et al. 2013).

3.3 Approaches to automatic subject indexing

As described in Section 3.1. above, methods to automatically index or classify are at its foundational level effectively the same—applying heuristic principles to computationally determine the subject of a document, and then assign an appropriate index term based on that. Approaches and differences between them may be grouped based on various criteria, and still the distinction will not always be clear-cut. The criteria followed here are based on the general context set out for this entry, that is, assigned subject indexing for purposes of information retrieval. The criteria are: a) application purposes; b) a more-or-less coherent body of published research following the approach; and, c) general approach: supervised learning, unsupervised learning, or string matching. The division that follows is also in line with previously published bibliometric analysis of the identified approaches (Golub and Larsen 2005) and a discussion on the same approaches as applied to web pages (Golub 2006b). Each approach is described via its definition, differences within the approach, application, and evaluation.

3.3.1 Text categorization

Text categorization or text classification are two terms that most often refer to automatic indexing of textual documents where both manually (intellectually) assigned documents and the target KOS exist. This is a machine-learning approach employing supervised learning whereby the algorithm "learns" about characteristics of target index terms based on characteristics of documents that had been manually pre-assigned those index terms. One of commonly used characteristics is word frequency; for example, words that often occur in documents assigned to the same index term as opposed to those that occur in documents assigned to other index terms.

The process comprises three major steps. First, a collection of documents manually (intellectually) indexed using a pre-defined KOS is chosen or created for the text categorization process. The documents in this collection are called training documents. In the second step, for each category a classifier is built, most often using the vector-space model. The classifiers are tested with a new set of documents from the collection; these are called test documents. Finally, the third step is the actual categorization where the classifier is applied to new documents.

The literature reports on a range of different ways to build classifiers, for example support vector machines (SVM) (e.g., Lee et al. 2012), artificial neural networks (e.g., Ghiassi et al. 2012), random forest learning (Klassen and Paturi 2010), adaptive boosting (AdaBoost) (Freund and Schapire 1997), to name a few considered to be state-of-the-art today. For an overview of different classifiers, see Mitchell 1997; for comparisons between them, see, Yang (1999) and Sebastiani (2002). Also, two or more different classifiers and ways to build them can be combined to make a classification decision—these are known as classifier committees or metaclassifiers (e.g., Liere and Tadepalli 1998; Wan et al. 2012; Miao et al. 2012).

Text categorization approaches can be divided into hard and soft; in hard, a decision is made as to whether the document does or does not belong to a category; in soft, a ranked list of candidate categories is created for each document and one or more of the top-ranked are chosen as the appropriate categories (Sebastiani 2002). The soft approach better reflects reality (cf. Section 4 where aboutness is discussed).

Text categorization has been applied to KOSs that incorporate hierarchies of concepts, such as Wikipedia, Open Directory Project, and Yahoo's Directory (for an overview, see, e.g., Ceci and Malerba 2007 and a workshop by Kosmopoulos et al. 2010). When compared to a flat approach, many have reported that including features based on the hierarchy structure in the classifier improves classification accuracy (e.g., McCallum et al. 1998; Ruiz and Srinivasan 1999; Dumais and Chen 2000). Li, Yang, and Park (2012) combined text categorization algorithms with WordNet and an automatically-constructed thesaurus and gained high effectiveness as measured by precision, recall, and F-measures (see below). Maghsoodi and Homayounpour (2011) have extended the feature vector of the SVM classifier by Wikipedia concepts and gained improved results (for the Farsi language). This is in line with research in document classification (see Section 3.3) where other features from existing KOSs have been used to improve the algorithm results.

Examples of test collections specially designed for use in text categorization include Reuters newswire stories (e.g., Reuters-21578), OHSUMED with metadata from MEDLINE, and WebKB for web pages, to name a few. However, for many document collections, there will be no training documents available to train and test the classifier. If there are no resources or possibilities to create one manually, approaches like semi-supervised learning and unsupervised learning can be adopted instead. For an overview of semi-supervised learning, see Mladenić and Grobelnik (2014). Unsupervised learning is basically document clustering described in the following section.

Evaluation in text categorization is often conducted by comparison against pre-assigned categories in test collections created for that task. Evaluation generally excludes deeper considerations of contexts like real-life end-user tasks and information practices. Furthermore, problems of using existing test collections for text categorization have been reported. Yang (1999) claims that the most serious problem in text categorization evaluations is the lack of standard data collections and shows how different versions even of the same collection have a strong impact on the performance. This corresponds to the well-established knowledge from inter-indexer consistency studies that human indexing is very inconsistent, and that inconsistency is an inherent feature of indexing, rather than a sporadic anomaly. Therefore Hjørland (2018, Section 3.2) concluded: "That human indexing is sometimes taken as the golden standard to which computer-indexing is adjusted is of course problematic in the light of the large degree of inconsistency found in empirical investigations and the uncertainty about how indexing should be evaluated."

Comparison between automatically and manually assigned categories is calculated using performance measures such as precision and recall used in information retrieval evaluation (see, for example, Manning, Raghavan, and Schütze 2008, chapter 8). In information retrieval, precision is defined as the fraction of retrieved documents that are relevant to the query and recall as the fraction of documents relevant to the query that are successfully retrieved.

$$\text{Precision} = \frac{|number\ of\ relevant\ retreived\ documents|}{|number\ of\ retreived\ documents|}$$

$$\text{Recall} = \frac{|number\ of\ relevant\ retreived\ documents|}{|number\ of\ relevant\ documents|}$$

Translated to automatic subject indexing, recall is calculated as the number of correct automatically assigned index terms divided by the number of manually assigned index terms. Precision is the number of correct automatically assigned index terms divided by the number of all automatically assigned index terms.

$$\text{Precision} = \frac{|number\ of\ correct\ automatically\ assigned\ terms|}{|number\ of\ all\ automatically\ assigned\ index\ terms|}$$

$$\text{Recall} = \frac{|number\ of\ correct\ automatically\ assigned\ terms|}{|number\ of\ manually\ assigned\ index\ terms|}$$

Macroaveraging and microaveraging are then used to obtain average performance over all index terms. Other aspects of algorithm performance may be evaluated, such as the speed of computation across the different steps of the process. For a detailed overview of these and other evaluation measures in text categorization, see Sebastiani (2002, 32-39). For further information on text categorization in terms of technical detail please refer to Sebastiani (2002) and for a more general overview to Mladenić and Grobelnik (2014).

3.3.2 Document clustering

Document clustering is the term most often used to refer to automatic construction of groups of topically related documents and automatic derivation of names for those groups of documents. Also, relationships between the groups of documents may be automatically determined, such as those that are hierarchical. No training documents are used from which the algorithm can "learn" to assign similar documents to the same topics. Therefore, this approach is known as unsupervised learning whereby the algorithm learns from existing examples without any "supervision."

Document clustering approach is best suited for situations when there is no target KOS at hand and no training documents, but the documents need to be topically grouped. It traditionally has been used to improve information retrieval, for example, when grouping search engine results into topics. On the other hand, automatic derivation of names and relationships is still a very challenging aspect of document clustering. "Automatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand" (Chen and Dumais 2000). Further, the clusters and relationships between them change as new documents are added to the collection; frequent changes of cluster names and relationships between them may not be user-friendly, for example, when applied for hierarchical topical browsing of a document collection. Koch, Zettergren and Day (1999) suggest that document clustering is better suited for organizing web search engine results.

The process of document clustering normally involves two major steps. First, documents in the collection at hand are typically each represented by vectors. The vectors are then compared to one another using vector similarity measures such as the cosine measure. A variety of heuristic principles may be applied when deriving vectors, as outlined in Section 3.1. Second, the chosen clustering algorithm is applied to group similar documents, name the clusters, and, if decided, derive relationships between clusters.

Similar to text categorization, there are two different approaches to clustering, hard and fuzzy (or soft). In hard clustering, one document may be a member of one cluster only, while in fuzzy clustering, any document may belong to any number of clusters. Hard clustering is the most

common approach to clustering. Its subtypes are partitional (also called flat) and hierarchical clustering. A typical example of partitional clustering is the k-means algorithm whereby the first step is to randomly create a *k* number of clusters and then new documents are added to the different clusters based on their similarity. As the document is added to the cluster, the clusters and their centroids (centre of a cluster) are re-computed. In hierarchical clustering, there are divisive and agglomerative algorithms. Divisive hierarchical clustering is a top-down approach in which, at start, all documents are grouped into one cluster that is then subdivided into smaller and smaller clusters up until each cluster comprises one document. Agglomerative hierarchical clustering is a bottom-up approach, starting from a set of clusters each comprising a single document, and gradually merging those with most similar vectors. Examples of less common approaches to clustering include self-organizing maps (see, e.g., Paukkeria et al. 2012; Lin, Brusilovsky, and He 2011; Saarikoski 2011), and genetic algorithms (see, e.g., Song, Yang, and Park 2011).

Bibliometrics also applies document clustering; to map research fields or represent subject categories. It does so by linking documents through establishing relations between documents that cite each other (co-citation), or that share same references sets (bibliographic coupling), for example. The underlying assumption is that the more connections are established, the more the documents have in common scientifically, which can also be interpreted as different research specializations, research areas or subject categories. In order to assign topical words to clusters instead of author or journal names from references, co-word analysis of titles, keywords or abstracts may be performed. Combining reference/citation analysis with co-word analysis is another approach. For more detail on these matters, see Åström (2014).

Evaluation in document clustering is often conducted by comparison to an existing manually created KOS or manually pre-assigned classes. Measures used include the number of correct decisions compared to all decisions (Rand index); precision, recall, and related. These are called external validity measures. There are also internal validity measures that estimate compactness, i.e., how close the documents are to each other in each cluster (the closer the better as this indicates better similarity), and separability, i.e., how distant two clusters are from one another (the more distant the better) (Frommholz and Abbasi 2014).

For further detail on similarity measures and other aspects of document clustering, please see chapters 16 and 17 of Manning, Raghvan, and Schütze (2008) and Frommholz and Abbasi (2014).

3.3.3 Document classification

A perhaps less established approach that we identify in this entry is that which tends to arise more specifically from the library and information science community whereby the purpose is to apply quality-controlled KOSs more directly to typical subject indexing (including classification) tasks in library catalogues or closely related information retrieval systems, in order to improve searching and browsing. For the purposes of this work and to distinguish between the previous two approaches, as well as to follow the line of previously published research (cf. Golub 2006a), we name this approach document classification. However, because this approach seems less established than the previous two, the community around it being less coherent, principles and methods applied may not be as homogeneous.

Apart from using quality-controlled KOSs for subject indexing and classification, this seems to be the only approach using string-matching between terms from the documents to be indexed and target index terms. As in text categorization and document clustering, the pre- processing of documents to be classified typically 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; while vector representations and manipulations are not necessary. Furthermore, examples using machine learning exist as seen from below. However, as to supervised machine learning, research points to scenarios where it may not work due the lack of training documents, especially for large KOSs; Wang (2009) and Waltinger et al. (2011), argue that *Dewey Decimal Classification*'s deep and detailed hierarchies lead to data sparseness and thus skewed distribution in supervised machine learning approaches.

While this approach is obviously different from document clustering in that here we have a target KOS, it shares this particular feature with the text categorization approach. Following the criteria to distinguish between approaches set out at the start of Section 3.3., the document classification approach is different from text categorization in that:

– its application tends to be tightly related to applying quality-controlled KOSs directly to typical subject indexing and classification tasks in library catalogues or related operative information retrieval systems;
– this seems to be the only approach using string-matching between terms from the documents to be indexed and target index terms, although examples using machine learning also exist, the latter being problematic due to training data sparseness especially for large KOSs.

However, like in many classifications, there are grey zones, which are discussed below.

Often the focus of research are publicly available, operative information systems using well-known KOSs. Examples include universal KOSs: *Dewey Decimal Classification* (*DDC*); Universal Decimal Classification (UDC); Library of Congress *Classification* (LC*C*); FAST (Faceted Application of Subject Terminology); German subject headings (Schlagwortnormdatei (SWD)); as well as subject-specific systems: *Medical Subject Headings* (*MeSH*), National Library of Medicine (NLM) classification system, *Engineering Index* classification system and thesaurus (used by the Compendex database), Inspec classification system and thesaurus, Fachinformationszentrums Technik (FIZ Technik) thesaurus and classification system, AGROVOC thesaurus, Humanities and Social Science Electronic Thesaurus (HASSET), and EuroVoc thesaurus. As the predicted relevance of this approach to the readers of the ISKO Encyclopedia is high, a more detailed, albeit non-exhaustive, overview of research will be provided in this section for illustration purposes. The overview is structured around the specific KOSs.

Online Computer Library Center's (OCLC) project Scorpion (OCLC Research 2004) built tools for automatic 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). Another OCLC project, WordSmith (Godby and Reighart 2001), was to develop software to extract significant noun phrases from a document. The idea behind it was that the precision of automatic indexing 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. Wolverhampton Web Library was a manually maintained library catalogue of British web resources, within which experiments on automating *DDC* classification were conducted (Jenkins et al. 1998). Resorting to already assigned *DDC*, Joorabchi and Mahdi (2011) extracted references from the document to be classified, compiled a list of publications that cite either the document to be classified or one of its references, and discovered their corresponding *DDC* numbers from existing library catalogues in order to then assign the most probable match to the document at hand. Similarly, Joorabchi and Mahdi (2013) assigned *DDC* and FAST by first identifying Wikipedia concepts in the document to be indexed/classified and then by searching WorldCat for records that contain those concepts. Then they compared the retrieved records against the document and assigned *DDC* and FAST to it from those with the highest matching score. Khoo et al. (2015) attempted to solve the problem of cross-searching unrelated libraries. To that extent, they created *DDC* terms and numbers from pre-existing Dublin Core metadata. The results indicate that best results are achieved when combined title, description, and subject terms are used. Further, they demonstrate how taking advantage of *DDC* hierarchies for disambiguation in simple string-matching can achieve results that are competitive with machine learning approaches, yet without the need for training documents.

In the Nordic WAIS/World Wide Web Project, 1993-1996 (Ardö et al. 1994; Koch 1994), automatic indexing of the World Wide Web and Wide Area Information Server (WAIS) databases using UDC was experimented with. A WAIS subject tree was built based on two top levels of UDC, i.e., fifty-one classes. UDC was also used by GERHARD, a robot-generated web index of web documents in Germany (Möller et al. 1999) that employed a multilingual version of UDC in English, German, and French.

Wartena and Sommer (2012) experimented with automatic indexing of articles in academic repositories using German subject headings (SWD). German subject headings have a thesaurus-like structure with synonyms, superordinate, and related terms. Also, about 40,000 terms have been enhanced with *DDC* classes. Like Khoo et al. (2015) (see above), they conclude that good results are achieved when applying string-matching, which they attribute to the enriched version of German subject headings. Junger (2014) reports on experiments run by the German National Library with the aim to use automatic indexing for online publications for which they have no resources to manually catalogue. They acquired commercial machine-learning software that has previously been specializing in automatic indexing of medical publications called Averbis. With catalogue librarians as evaluators, the recall was considered high but precision too low to be satisfactory, attributing this to lack of disambiguation mechanisms; they proposed co-occurrence analysis and related techniques to be implemented in the future.

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 at the time), where it was used to support hierarchical browsing.

One of the most well-researched automatic indexing software applications was created in 1996 by the National Library of Medicine, known as Medical Text Indexer (MTI) (a lot of publications and other resources about it can be found at its website, https://ii.nlm.nih.gov/Publications/). It is semi-automatic software aimed at assigning

MeSH. The general approach is combining the intellectual work built into the rich UMLS Metathesaurus (UMLS—Unified medical language system), extracted *MeSH* terms from related citations, with comprehensive indexing rules and machine learning. In one of the most recent articles titled "12 Years On—Is the NLM Medical Text Indexer Still Useful and Relevant?," Mork, Aronson, and Demner-Fushman (2017) show how indexers have continually increased their use of the MTI, from 15.75% of the articles indexed with it in 2002, to 62.44% in 2014, at the same time also spreading to new subject areas of use, indicating its usefulness. Furthermore, the MTI performance statistics show significant improvement in precision and F-measures while they point to the need to improve recall, too. One point for further research and development is to resort more to machine learning while keeping the existing components. Of other medical document types, Pratt (1997) experimented with organizing search results into *MeSH* categories. Lüschow and Wartena (2017) applied k-nearest-neighbour (kNN) algorithm to a collection of medical documents with pre-assigned classes from several classification systems, with the aim of using them as a basis on which to automatically assign the National Library of Medicine classification system, thus using already assigned classes from other classification systems instead of using, e.g., book titles or keywords as the content representation for each document.

"All" Engineering was a robot-generated web index of about 300,000 web documents, developed as an experimental module of the manually created subject gateway Engineering Electronic Library (EELS) (Koch and Ardö 2000). *Engineering Index* (*Ei*) thesaurus was used; in this thesaurus, terms are enriched with their mappings to *Ei* classification scheme. The project proved the importance of applying a good KOS in achieving the automatic indexing accuracy: 60% of documents were correctly classified, using only a very simple string-matching algorithm based on a limited set of heuristics and simple weighting. Another robot-generated web index, Engine-e, used a slightly modified automatic indexing approach to the one developed in "All" Engineering (Lindholm, Schönthal, and Jansson 2003). Engine-e provided subject browsing of engineering documents based on *Ei* terms, with six broader categories as starting points. Golub, Hamon, and Ardö (2007) applied string-matching where the *Ei* thesaurus terms were enriched with automatically extracted terms from bibliographic records of the Compendex database, using multi-word morpho-syntactic analysis and synonym acquisition, based on the existing preferred and synonymous terms (as they gave best precision results). Golub (2011) worked with *Ei* to automatically organize web pages into hierarchical structures for subject browsing, achieving results suggesting how a KOS with a sufficient number of entry terms designating classes could significantly increase performance of automatic indexing algorithms. Further, if the same KOS had an appropriate hierarchical structure, it would provide a good browsing structure for the collection of automatically classified documents.

Plaunt and Norgard (1997) applied a supervised training algorithm based on extracting lexical terms from bibliographic records and associating them with manually-assigned INSPEC thesaurus terms. Project BINDEX (Bilingual Automatic Parallel Indexing and Classification) (Maas et al. 2002) applied automatic indexing of abstracts in engineering available in the English and German languages. It used the English Inspec thesaurus and classification system, as well as, FIZ Technik's bilingual thesaurus and classification system. Morpho-syntactic analysis of a document was performed. It involved identification of single and multiple-word terms, tagging and lemmatization, and homograph resolution. Keywords were extracted and matched against the thesauri, and then classification codes were derived. Keywords above a certain threshold which were not in the thesaurus were assigned as free index terms. Enriching records with other terms than from the KOS at hand might lead to improved retrieval. To that extent, Joorabchi and Mahdi (2014) experimented with adding Wikipedia concepts to existing library records.

Lauser and Hotho (2004) applied a support vector machines (SVM) algorithm to index a collection of agricultural documents with the AGROVOC thesaurus. The algorithm improved when they made use of the semantic information contained in AGROVOC. Similarly, Medelyan and Witten (2008) used KEA, a Naïve Bayes algorithm for extracting both derived and assigned index terms and achieved good performance with little training data, because they also made use of the AGROVOC semantic information.

Of other examples, De Campos and Romero (2008) used machine learning to classify parliamentary resolutions from the regional Parliament of Andalucía at Spain using EuroVoc. El-Haj et al. (2013) experimented with applying HASSET terms to the UK Data Archive/UK Data Service data-related document collection. Their approach was based on applying an open source, machine-learning keyphrase extractor KEA (Keyphrase Extraction Algorithm).

As we see from the examples above, in many of the cases, the relationships built into KOSs are explored with favorable results. Willis and Losee (2013) specifically experimented with just that. They employed four thesauri in order to determine to what degree the in-built relationships may be used to the advantage of automatic subject indexing. Their results indicate a great potential, albeit the degree of success seems to be dependent on the thesaurus as well as collection.

A major advantage of this approach is that it does not require training documents, while still maintaining a predefined structure of the KOS at hand. If using a high-quality KOS, e.g., a well-developed classification scheme, it will also be suitable for subject searching and browsing in information retrieval systems. Apart from improved information retrieval, another motivation to apply a KOS in automatic classification is to re-use the intellectual effort that has gone into creating such a KOS. It can be employed with vocabularies containing uneven hierarchies or sparse distribution across a given collection.

As for evaluation methods, measures such as precision and recall, and F-measure are commonly used. This seems to be the only approach where at least the discussion is occasionally brought up calling for the need to attend to the complexities of evaluation closer to real-life needs and scenarios. Even aspects such as automatic indexing warrant are taken on; Chung, Miksa, and Hastings (2010) conclude that literary warrant is more suited in automatic indexing of scientific articles than user warrant.

4.0 Application in operative systems

The discussion on how applicable automatic subject indexing is today calls for looking into at least several connected issues. Theoretically, 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 (Svenonius 2000, 46-49). In algorithms, inferences are made such as: if document A is on subject X, then if document B is sufficiently similar to document A (e.g., they share similar words or references), then document B is on that subject. Another critique given is the lack of theoretical justifications for vector manipulations, such as the cosine measure that is often used to obtain vector similarities (Salton 1991, 975). Further, it is assumed that concepts have names, which can be more common in, for example, natural sciences, but much less so in humanities and social sciences, although attempts to address this have been undertaken more recently (see Section 3.1).

A variety of factors contribute to the challenge of automatic subject indexing. Texts are a complex cognitive and social phenomenon, and cognitive understanding of text engages many knowledge sources, sustains multiple inferences, and involves a personal interpretation (Moens 2000, 7-10). Morris (2010) investigated individual differences in the interpretation of text meaning using lexical chains (groups of semantically related words) based on three texts and with twenty-six participants; the results showed about 40% difference in interpretation. Research in automatic understanding of text covers the linguistic coding (vocabulary, syntax, and semantics of the language and discourse properties), domain world knowledge, shared knowledge between the creator and user of the text, and the complete context of the understanding at a specific point in time including the ideology, norms, background of the user, and the purposes of using the text. In 2003, Lancaster claimed that existing automatic subject indexing tools are far from being able to handle the complexities, and in applications it is seldom possible to go much further beyond vocabulary and syntax analysis. The difficulty of dealing with semantics and more advanced levels is reflected in the fact that the methods used today are not particularly new although not as rudimentary when first used (Lancaster 2003, 330-331).

Still, software vendors and experimental researchers speak of the high potential of automatic indexing tools. While some claim to entirely replace manual indexing in certain subject areas (e.g., Roitblat, Kershaw and Oot 2010), others recognize the need for both manual (human) and computer-assisted indexing, each with its advantages and disadvantages (e.g., Anderson and Perez-Carballo 2001; Svarre and Lykke 2014). Reported examples of operational information systems where machine-aided indexing is applied include NASA's MAI software, which was shown to increase production and improve indexing quality (Silvester 1997), and the Medical Text Indexer at the US National Library of Medicine, which, by 2017, was consulted by indexers in over 60% or articles indexing (Mork, Aronson and Demner-Fushman 2017).

However, hard evidence on the success of automatic indexing tools in operating information environments is scarce; research is usually conducted in laboratory conditions, excluding the complexities of real-life systems and situations. The practical value of automatic indexing tools is largely unknown due to problematic evaluation approaches. Having reviewed a large number of automatic indexing studies, Lancaster concluded that the research comparing automatic versus manual indexing is flawed (2003, 334). One common evaluation approach is testing the quality of retrieval based on the assigned index terms. But retrieval testing is fraught with problems, too; the results depend on many factors, so retrieval testing cannot isolate the quality of the index terms. Another approach is to measure indexing quality directly. One method of doing so is to compare automatically assigned metadata terms against existing human-assigned terms or classes of the document collection used (as a "gold standard"), but this method also has problems. When indexing, people make errors, such as related to exhaustivity (too many or too few subjects assigned) or specificity (usually because the assigned subject is not the most specific available); they may omit important subjects or assign an obviously incorrect subject (see also Hjørland 2017 for a detailed discussion on different aspects of aboutness). In addition, it has been

reported that different people, whether users or professional subject indexers, assign different subjects to the same document. One reason for this are differences in the approach: on the one hand, following the rationalist idea that there is one correct way to index a document (or a collection), and on the other one, the pragmatic idea that different purposes and users may need different indexing (Hjørland 2018). Therefore, existing metadata records cannot be used as "the gold standard;" the classes assigned by algorithms (but not human-assigned) might be wrong or might be correct but omitted during human indexing by mistake or by abiding to a certain indexing policy.

In order to address the complexities surrounding the problem of aboutness, Golub et al. (2016) propose a comprehensive framework involving three major steps: evaluating indexing quality directly through assessment by an evaluator or through comparison with a gold standard, evaluating the quality of computer-assisted indexing directly in the context of an indexing workflow, and evaluating indexing quality indirectly through analyzing retrieval performance. The framework still needs to be tested empirically, and it is expected that much more research is required to develop appropriate evaluation designs for such complex phenomena involving subject indexing and retrieval and information interaction in general.

While evaluation approaches often assume that human indexing is best, and that the task of automatic indexing is to meet the standards of human indexers, more serious scholarship needs to be devoted to evaluation in order to further our understanding of the value of automatic subject assignment tools and to enable us to provide a fully informed input for their development and enhancement. Hjørland (2011) points to the problematics of evaluating indexing on an example of an empirical study and discusses this through a theory of knowledge point of view, while analyzing its epistemological position. He concludes by proposing that the ideal formula for the future of indexing is that the human indexer takes what automatic indexing is good at (once this is understood) and invest their resources on the value-added indexing that requires human judgment and interpretation. This may be in line with machine-aided indexing in operative systems like Medical Text Indexer mentioned at the start of this section.

5.0 Conclusions

Basic principles applied in various approaches to automatically assign index terms are at its foundational level effectively the same. The focus is still largely at the level of words rather than concepts and commonly includes punctuation and stop-word removal, stemming, heuristic rules, and vector representations and manipulations. While attempts to determine concepts rather than words exist and include LSI and word2vec as well as exploiting relationships from existing KOSs, much more research is needed in this respect.

Approaches to automatic subject indexing may be grouped based on various criteria; those followed in this work are based on the general context set out for this entry, that is assigned subject indexing for purposes of information retrieval. The named approaches are also in line with previous research and include: text categorization, document clustering and document classification. Major differences between them include application purposes and presence or absence of machine learning, as well as whether machine learning is supervised or unsupervised. The document classification approach employs, more than others, subject indexing languages such as classification schemes, subject headings systems, and thesauri, which are also suitable for subject searching and browsing in an information retrieval system (although often suggested improvements such as being more up-to-date, end-user friendly, etc. should be addressed). Not the least, exploiting the intellectual work that has been invested into creating such subject indexing languages in order to improve automatic indexing has shown to be a worthwhile path to explore more extensively in the future.

Due to complexities of aboutness, existing experimental systems and approaches have not been adequately tested and therefore knowledge about their usefulness for operational systems seems to be flawed. A recently proposed comprehensive evaluation framework involves three major steps: evaluating indexing quality directly through assessment by an evaluator or through comparison with a gold standard, evaluating the quality of computer-assisted indexing directly in the context of an indexing workflow and evaluating indexing quality indirectly through analyzing retrieval performance. Further research is needed to empirically test it as well as devise most appropriate evaluation approaches for different specific contexts.

References

Anderson, James D. and Jose Perez-Carballo. 2001. "The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. Part II: Machine Indexing, and the Allocation of Human versus Machine Effort." *Information Processing and Management* 37, no. 2: 255-77.

Ardö, A. et al. 1994. "Improving Resource Discovery and Retrieval on the Internet: The Nordic WAIS/World Wide Web Project Summary Report." *NORDINFO Nytt* 17, no. 4: 13-28.

Åström, Fredrik. 2014. "Bibliometrics and Subject Representation." In *Subject Access to Information: An Interdisciplinary Approach*, ed. Koraljka Golub, 107-17. Santa Barbara: Libraries Unlimited.

Baxendale, Phyllis B. 1958. "Machine-made Index for Technical Literature—An Experiment." *IBM Journal of Research and Development* 2: 354-61.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993-1022.

Bradshaw, Shannon and Kristian Hammond. 1999. "Constructing Indices from Citations in Collections of Research Papers." In *Knowledge, Creation, Organization and Use: ASIS '99: 62nd ASIS Annual Meeting, Washington, DC, October 31-November 4, 1999,* ed. Marjorie K. Hlava and Larry Woods. Silver Spring, MD: American Society for Information Science, 741-50.

Ceci, Michelangelo and Donato Malerba. 2003. "Hierarchical Classification of HTML Documents with WebClassII." In *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14-16, 2003,* ed. Fabrizio Sebastiani. Berlin: Springer, 57-72. doi:10.1007/3-540-36618-0_5

Chen, Hao and Susan Dumais. 2000. "Bringing Order to the Web: Automatically Categorizing Search Results." In *Proceedings of the ACM CHI 2000 Human Factors in Computing Systems Conference, The Hague, The Netherlands, April 1-6, 2000,* ed. Thea Turner, Gerd Szwillus, Mary Czerwinski, Fabio Peterno, and Steven Pemberton. New York: ACM Press, 145-52.

Chung, EunKyung, Shawne Miksa and Samantha K. Hastings. 2010. "A Framework of Automatic Subject Term Assignment for Text Categorization: An Indexing Conception-based Approach." *Journal of the American Society for Information Science and Technology* 61: 688-99.

De Campos, Louis M. and Alfonso E. Romero. 2009. "Bayesian Network Models for Hierarchical Text Classification from a Thesaurus." *International Journal of Approximate Reasoning* 50: 932-44.

Deerwester, Scoot, Susan T. Dumais, Thomas K. Landauer, George W. Furnas and Louis Beck. 1988. "Improving Information Retrieval with Latent Semantic Indexing." In *ASIS '88: Proceedings of the 51st ASIS Annual Meeting, Atlanta, Georgia, October 23-27, 1988,* ed. Christine L. Borgman, and Edward Y. H. Pai. Proceedings of the ASIS Annual Meeting 25. Medford: Learned Information, 36-40.

Dehghani Mostafa, Hosein Azarbonyad, Maarten Marx and Jaap Kamps. 2015. "Sources of Evidence for Automatic Indexing of Political Texts." In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015; Proceedings*, ed. Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr. Lecture Notes in Computer Science 9022. Cham: Springer, 568-73. doi:10.1007/978-3-319-16354-3_63

Driscoll, James R. 1991. "The Operation and Performance of an Artificially Intelligent Keywording System." *Information Processing & Management 27*: no. 1: 43-54.

Dumais, Susan T. and Hao Chen. 2000. "Hierarchical Classification of Web Content." In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece,* ed. Emmanuel Yannakoudakis, Nicholas J. Belkin, Mun-Kew Leong, and Peter Ingwersen. [S. l.]: ACM 2000, 256-63.

El-Haj, Mahmoud, Lorna Balkan, Suzanne Barbalet, Lucy Bell, and John Shepherdson. "An Experiment in Automatic Indexing Using the HASSET Thesaurus." In *2013 5th Computer science and Electronic Engineering Conference (CEEC), 17th-18th September 2013, University of Essex, UK: Conference Proceedings.* [Piscataway, NJ]: IEEE, 13-8. doi:10.1109/CEEC.2013.6659437

Esser, Daniel, Daniel Schuster, Klemens Muthmann, Michael Berger and Alexander Schill. 2012. "Automatic Indexing of Scanned Documents: A Layout-based Approach." In *Document Recognition and Retrieval XIX: Part of the IS&T/SPIE 24th Annual Symposium on Electronic Imaging, 22-26 January 2012, San Francisco, CA USA,* ed. Christian Viard-Gaudin, and Richard Zanibbi. SPIE Proceedings 8297. Washington: SPIE. doi:10.1117/12.908542

Flett, Alan and Stuart Laurie. 2012. "Applying Taxonomies through Auto-Classification." *Business Information Review* 29, no. 2: 111-20.

Freund, Yoav and Robert E. Schapire. 1997. "A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting." *Journal of Computer and System Sciences* 55: 119-39.

Frank, Eibe and Gordon W. Paynter. 2004. "Predicting Library of Congress Classifications from Library of Congress Subject Headings." *Journal of the American Society for Information Science and Technology* 55: 214-27.

Frommholz, Ingo and Muhammad Kamran Abbasi. 2014. "Automated Text Categorization and Clustering." In *Subject Access to Information: An Interdisciplinary Approach*, ed. Koraljka Golub. Santa Barbara: Libraries Unlimited, 117-31.

Fuhr, Norbert and Gerhard Knorz. 1984. "Retrieval Test Evaluation of a Rule-based Automated Indexing (AIR/PHYS)." In *Research and Development in Information retrieval: Proceedings of the Third Joint BCS and ACM Symposium, King's College, Cambridge, 2-6 July 1984,* ed. Cornelis Joost van Rijsbergen. Cambridge: Cambridge University Press, 391-408.

Ghiassi, Manoochehr, Michael Olschimke, Brian Moon and Paul Arnaudo. 2012. "Automated Text Classification Using a Dynamic Artificial Neural Network Model." *Expert Systems with Applications* 39: 10967-76.

Gil-Leiva, Isidoro. 2017. "SISA – Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules Versus TF-IDF Rules." *Knowledge Organization* 44: 139-62.

Godby, C. Jean and Ray R. Reighart. 2001. "The WordSmith Indexing System." *Journal of Library Administration* 34, nos. 3-4: 375-85.

Goldberg, Yoav and Omer Levy. 2014. "Word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method." https://arxiv.org/abs/1402.3722

Golub, Koraljka. 2006a. "Automated Subject Classification of Textual Web Documents." *Journal of Documentation* 62: 350-71.

Golub, Koraljka. 2006b. "Automated Subject Classification of Textual Web Pages, Based on a Controlled Vocabulary: Challenges and Recommendations." *New Review of Hypermedia and Multimedia* 12, no. 1: 11-27.

Golub, Koraljka. 2016. "Potential and Challenges of Subject Access in Libraries Today on the Example of Swedish Libraries." *International Information & Library Review* 48: 204-10.

Golub, Koraljka and Anders Ardö. 2005. "Importance of HTML Structural Elements and Metadata in Automated Subject Classification." In *Research and Advanced Technology for Digital Libraries: 9th European Conference, ECDL 2005, Vienna, Austria, September 18-23, 2005; Proceedings,* ed. Andreas Rauber, Stavros Christodoulakis and Min A. Tjoa. Berlin: Springer, 368-78.

Golub, Koraljka, Thierry Hamon and Anders Ardö. 2007. "Automated Classification of Textual Documents based on a Controlled Vocabulary in Engineering." *Knowledge Organization* 34: 247-63.

Golub, Koraljka and Birger Larsen. 2005. "Different Approaches to Automated Classification: Is There an Exchange of Ideas?" In *Proceedings of ISSI 2005: The 10th International Conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, July 24-28, 2005*, ed. Peter Ingwersen and Birger Larsen. Stockholm: Karolinska University Press, 270-4.

Golub, Koraljka, Dagobert Soergel, George Buchanan, Douglas Tudhope, Marianne Lykke, and Debra Hiom. 2016. "A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval." *Journal of the Association for Information Science and Technology* 67: 3-16.

Grobelnik, Marko and Dunja Mladenić. 2005. "Simple Classification into Large Topic Ontology of Web Documents." *Journal of Computing and Information Technology* 13: 279-85.

Gövert, Norbert, Mounia Lalmas, and Norbert Fuhr. 1999. "A Probabilistic Description-oriented Approach for Categorising Web Documents." In *CIKM99: Eighth International Conference on Information and Knowledge Management, Kansas City, MO, USA – November 02 - 06, 1999*, ed. Susan Gauch. New York, NY: Association for Computing Machinery, 475-82.

Hamm, Sandra and Kurt Schneider. 2015. "Automatische Erschliessung von Universitätsdissertationen." *Dialog mit Bibliotheken* 27, no. 1: 18-22.

Harris, Zellig S. 1954. "Distributional Structure." *Word* 10, no. 23: 146-62.

Hartigan, John A. 1996. "Introduction." In *Clustering and Classification,* edited by Phipps Arabie, Lawrence J. Hubert and Geert de Soete. Singapore: World Scientific, 3-5.

Hjørland, Birger. 2011. "The Importance of Theories of Knowledge: Indexing and Information Retrieval as an Example." *Journal of the American Society for Information Science and Technology* 62: 72–7.

Hjørland, Birger. 2017. "Subject (of Documents)." *Knowledge Organization* 44: 55-64.

Hjørland, Birger. 2018. "Indexing: Concepts and Theory." *Knowledge Organization* 45: 609-39.

Hlava, Majorie K. 2009. "Understanding 'Rule Based' vs. 'Statistics Based' Indexing Systems: Data Harmony White Paper." Reprinted from *Information Outlook* with permission, updated April, 2009. https://web.archive.org/web/20090417210346/http://www.dataharmony.com:80/library/whitePapers/auto_indexing_rule-based_vs_statistics-based.htm

Hofmann, Thomas. 2001. "Unsupervised Learning by Probabilistic Latent Semantic Analysis." *Machine Learning* 42, no. 1: 177-96.

Hu, Yi and Wenjie Li. 2011. "Document Sentiment Classification by Exploring Description Model of Topical Terms." *Computer Speech & Language* 25: 386-403.

Huang, Lan, David Milne, Frank Eibe, and Ian H. Witten. 2012. "Learning a Concept-based Document Similarity Measure: Report." *Journal of the American Society for Information Science and Technology* 63: 1593-1608.

Humphrey, Susanne M., Aurélie Névéol, Allen Browne, Julien Gobeil, Patrick Ruch and Stéfan J. Darmoni. 2009. "Comparing a Rule-Based Versus Statistical System for Automatic Categorization of MEDLINE Documents According to Biomedical Specialty." *Journal of the American Society for Information Science and Technology* 60: 2530-9.

Hwang, San-Yih, Wan-Shiou Yang and Kang-Di Ting. 2010. "Automatic Index Construction for Multimedia Digital Libraries." *Information Processing & Management* 46: 295-307.

International Organization for Standardization. 1985. *Documentation, Methods for Examining Documents, determining their Subjects, and Selecting Index Terms.* International Standard ISO 5963. [Geneva]: International Organization for Standardization.

Jenkins, Charlotte, Mike Jackson, Peter Burden and Jon Wallis. 1998. "Automatic Classification of Web Resources Using Java and Dewey Decimal Classification." *Computer Networks & ISDN Systems* 30: 646-8.

Jones, Kevin P. and Colin L. M. Bell. 1992. "Artificial Intelligence Program for Indexing Automatically (AIPIA)." In *Online Information* 92*: 16th International Online Information Meeting Proceedings, London, 8-10 December 1992*, ed. David I. Raitt. Medford, NJ: Learned Information, 187-96.

Joorabchi, Arash and Abdulhussain E. Mahdi. 2014. "Towards Linking Libraries and Wikipedia: Automatic Subject Indexing of Library Records with Wikipedia Concepts." *Journal of Information Science* 40: 211-21.

Joorabchi, Arash and Abdulhussain E. Mahdi. 2013. "Classification of Scientific Publications According to Library Controlled Vocabularies: A New Concept Matching-based Approach." *Library Hi Tech* 31: 725-47.

Junger, Ulrike. 2014. "Can Indexing Be Automated? The Example of the Deutsche Nationalbibliothek." *Cataloging & Classification Quarterly* 52, no. 1: 102-9.

Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.* Cambridge, MA: MIT Press.

Keyser, Pierre de. 2012. *Indexing: From Thesauri to the Semantic Web.* Oxford: Chandos.

Khoo, Michael John, Jae-wook Ahn, Ceri Binding, Hilary Jane Jones, Xia Lin, Diana Massam and Douglas Tudhope. 2015. "Augmenting Dublin Core Digital Library Metadata with Dewey Decimal Classification." *Journal of Documentation* 71: 976-98.

Klassen, Myungsook and Nikhila Paturi. 2010. "Web Document Classification by Keywords Using Random Forests." *Communications in Computer and Information Science* 88, no. 2: 256-61.

Koch, Traugott. 1994. "Experiments with Automatic Classification of WAIS Databases and Indexing of WWW." In *Internet World & Document Delivery World International 94, held in London in May 1994: Proceedings of the Second Annual Conference,* ed. John W. T. Smith. London: Mecklermedia, 112-5.

Koch, Traugott and Anders Ardö. 2000. "Automatic Classification." https://web.archive.org/web/20050301133443/http://www.lub.lu.se:80/desire/DESIRE36a-overview.html

Koch, Traugott, Ann-Sofie Zettergren and Michael Day. 1999. "Provide Browsing Using Classification Schemes." https://web.archive.org/web/20050403233258/http://www.lub.lu.se/desire/handbook/class.html

Kosmopoulos, Aris, Eric Gaussier, Georgios Paliouras and Sujeevan Aseervatham. 2010. "The ECIR 2010 Large Scale Hierarchical Classification Workshop." *ACM SIGIR Forum* 44, no. 1: 23-32.

Lancaster, Frederick W. 2003. *Indexing and Abstracting in Theory and Practice.* 3rd ed. London: Facet.

Larson, Ray R. 1992. "Experiments in Automatic Library of Congress Classification." *Journal of the American Society for Information Science* 432: 130-48.

Lauser, Boris and Andreas Hotho. 2003. "Automatic Multi-label Subject Indexing in a Multilingual Environment." In *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003; Proceedings,* ed. Traugott Koch and Ingeborg T. Sølvberg. Lecture Notes in Computer Science 2769. Berlin: Springer, 140-51.

Lee, Lam, Hong Wan, Chin Rajkumar, and Heng Isa. 2012. "An Enhanced Support Vector Machine Classification Framework by Using Euclidean Distance Function for Text Document Categorization." *Applied Intelligence* 37: 80-99.

Liere, Ray and Prasad Tadepalli. 1998. "Active Learning with Committees: Preliminary Results in Comparing Winnow and Perceptron in Text Categorization." In *Conference on Automated Learning and Discovery, June 11-13, 1998, Carnegie Mellon University, Pittsburgh, PA.* S.l.: s.n., 591-6.

Lin, Yi-ling, Peter Brusilovsky and Daqing He. 2011. "Improving Self-organising Information Maps as Navigational Tools: A Semantic Approach." *Online Information Review* 353: 401-24.

Lindholm, Jessica, Tomas Schönthal and Kjell Jansson. 2003. "Experiences of Harvesting Web Resources in Engineering using Automatic Classification." *Ariadne* no. 37. http://www.ariadne.ac.uk/issue37/lindholm/

Liu, Rey-Long. 2010. "Context-based Term Frequency Assessment for Text Classification." *Journal of the American Society for Information Science and Technology* 61: 300-9.

Lösch, Mathias, Ulli Waltinger, Wolfram Hortsmann and Alexander Mehler. 2011. "Building a DDC-annotated Corpus from OAI metadata." *Journal of Digital Information* 12, no. 2. https://journals.tdl.org/jodi/index.php/jodi/article/view/1765

Luhn, Hans P. 1957. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information." *IBM Journal of Research and Development* 1: 309-17.

Lüschow, Andreas and Christian Wartena. 2017. "Classifying Medical Literature Using k-Nearest-Neighbours Algorithm." In *NKOS 2017, 17th European Networked Knowledge Organization Systems (NKOS) Workshop, co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017), Thessaloniki, Greece, September 21st, 2017; Proceedings,* ed. Philipp Mayr, Douglas Tudhope, Koraljka Golub, Christian Wartena and Ernesto William De Luca. http://ceur-ws.org/Vol-1937/paper3.pdf

Maghsoodi, Nooshin and Mohammad Mehdi Homayounpour. 2011. "Using Thesaurus to Improve Multiclass

Text Classification." In *Computational Linguistics and Intelligent Text Processing: CICLing 2011*, ed. Alexander F. Gelbukh. Lecture Notes in Computer Science 6609. Berlin: Springer, 244-53.

Mahfouz, Tarek. 2011. "Unstructured Construction Document Classification Model through Support Vector Machine (SVM)." In *Computing in Civil Engineering: Proceedings of the 2011 ASCE International Workshop on Computing in Civil Engineering, June 19-22, Miami, Florida,* ed. Yimin Zhu and Raymond R. Issa. Reston, VA.: American Society of Civil Engineers, 126-33.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* New York: Cambridge University Press.

Maron, Melvil E. 1961. "Automatic Indexing: An Experimental Inquiry." *Journal of the Association for Computing Machinery* 8: 404-17.

Martinez-Alvarez, Miguel, Sirvan Yahyaei and Thomas Roelleke. 2012. "Semi-automatic Document Classification: Exploiting Document Difficulty." In *Advances in Information Retrieval: 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012; Proceedings* ed. Richard Baeza-Yates. Lecture Notes in Computer Science 7224. Berlin: Springer, 468-71.

Mazzocchi, Fulvio. 2018. "Knowledge Organization System (KOS): An Introductory Critical Account." *Knowledge Organization* 45: 54-78.

McCallum, Andrew, Ronald Rosenfeld, Tom Mitchell and Andrew Y. Ng. 1998. "Improving Text Classification by Shrinkage in a Hierarchy of Classes." In *Machine Learning: Proceedings of the Fifteenth International Conference (ICML '98), Madison, Wisconsin, July 24-27, 1998,* ed. Jude W. Shavlik. San Francisco, CA: Morgan Kaufmann, 359-67.

Medelyan, Olena and Ian H. Witten. 2008. Domain-Independent Automatic Keyphrase Indexing with Small Training Sets. *Journal of the American Society for Information Science and Technology* 59: 1026-40.

Meng, Jiana, Hongfei Lin and Yuhai Yu. "A Two-stage Feature Selection Method for Text Categorization." *Computers and Mathematics with Applications* 62, no. 7: 2793-800.

Miao, Duoqian, Qiguo Duan, Hongyun Zhang and Na Jiao. 2009. "Rough Set Based Hybrid Algorithm for Text Classification." *Expert Systems with Applications* 36: 9168-74.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *The Computing Research Repository (CoRR),* January 2013. https://arxiv.org/abs/1301.3781

Mladenić, Dunja and Marko Grobelnik. 2014. "Machine Learning on Text." In *Subject Access to Information: An Interdisciplinary Approach,* ed. Koraljka Golub, 132-8. Santa Barbara, CA: Libraries Unlimited.

Moens, Marie-Francine. 2000. *Automatic Indexing and Abstracting of Document Texts*. Boston: Kluwer.

Mork, James, Alan Aronson and Dina Demner-Fushman. 2017. "12 Years on – Is the NLM Medical Text Indexer Still Useful and Relevant?" *Journal of Biomedical Semantics* 8, no. 8. doi:10.1186/s13326-017-0113-5

Morris, Jane. 2010. "Individual Differences in the Interpretation of Text: Implications for Information Science." *Journal of the American Society for Information Science and Technology* 61: 141-9.

Mynarz, Jindřich and Ctibor Škuta. 2010. "Integration of Automatic Indexing System within the Document Flow in Grey Literature Repository." In: *Twelfth International Conference on Grey Literature: Transparency in Grey Literature, 6-7 December 2010, ed.* Dominic J. Farace and Jerry Frantzen. Amsterdam: TextRelease. http://www.greynet.org/images/GL12_S3P,_Mynarz_and_Skuta.pdf

Möller, Gerhard, Kai-Uwe Carstensen, Bernd Diekman, and Han Wätjen. 1999. "Automatic Classification of the WWW Using the Universal Decimal Classification." In *Online Information 99: Proceedings, London, 7-9 December 1999,* ed. David Raitt. Oxford: Learned Information Europe, 231-8.

Mu, Jin, Karsten Stegmann, Elijah Mayfield, Carolyn Rose, and Frank Fischer. 2012. "The ACODEA Framework: Developing Segmentation and Classification Schemes for Fully Automatic Analysis of Online Discussions." *International Journal of Computer-Supported Collaborative Learning* 7: 285-305.

Maas, Dieter, Rita Nuebel, Catherine Pease, and Paul Schmidt. 2002. "Bilingual Indexing for Information Retrieval with AUTINDEX." In *LREC 2002: Third International Conference on Language Resources and Evaluation, 29th, 30th & 31st May, Las Palmas de Gran Canaria (Spain); proceedings,* ed. Manuel González Rodríguez and Carmen Paz Suárez Araujo. Paris: European Language Resources Association, 1136-49.

National Library of Medicine. 2016. "NLM Medical Text Indexer (MTI)." https://ii.nlm.nih.gov/MTI/

OCLC Research. 2004. *Scorpion.* http://www.oclc.org/research/software/scorpion/default.htm

Page, Larry, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. *The Pagerank Citation Ranking: Bringing Order to the Web.* http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Paukkeria, Mari-Sanna, Alberto Pérez García-Plazab, Víctor Fresnob, Raquel Martínez Unanueb, and Timo-Honkela. 2012. "Learning a Taxonomy from a Set of Text Documents." *Applied Soft Computing* 12: 1138-48.

Perry, James W., Allen Kent, and Madeline M. Berry. 1955. "Machine Literature Searching X: Machine Language; Factors Underlying its Design and Development." *American Documentation* 6: 242.

Plaunt, Christian and Barbara A. Norgard. 1998. "An Association-based Method for Automatic Indexing with a Controlled Vocabulary." *Journal of the American Society for Information Science* 49: 888-902.

Pratt, Wanda. 1997. "Dynamic Organization of Search Results Using the UMLS." In *The Emergence of 'Internetable' Health Care: Systems that Really Work; 1997 AMIA Annual Fall Symposium, formerly SCAMC; A Conference of the American Medical Informatics Association, October 25-29, 1997, Opryland Hotel, Nashville, TN.; Proceedings,* ed. Daniel R. Masys. Philadelphia: Hanley & Belfus: 480-4.

Rasmussen Neal, D., ed. 2012. *Indexing and Retrieval of Non-text Information*. Berlin: De Gruyter Saur.

Roelleke, Thomas. 2013. *Information Retrieval Models: Foundations and Relationships.* San Rafael, CA: Morgan & Claypool.

Roitblat, Herbert L., Anne Kershaw, and Patrick Oot. 2010. "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review." *Journal of the American Society for Information Science and Technology* 61: 70-80.

Ruiz, Miguel E. and Padmini Srinivasan. 1999. "Hierarchical Neural Networks for Text Categorization." In *Proceedings of SIGIR '99: 22nd International Conference on Research and Development in Information Retrieval, University of California, Berkeley, August 1999;* ed. Marti Hearst, Fredric C. Gey, and Richard Tong. Association for Computing Machinery. New York, N.Y.: Association for Computing Machinery, 281-2.

Ruiz, Miguel E., Alan R. Aronson and Marjorie Hlava. 2008. "Adoption and Evaluation Issues of Automatic and Computer Aided Indexing Systems." In *ASIST 2008: Proceedings of the 71st ASIS&T Annual Meeting: People Transforming Information - Information Transforming People,* ed. Andrew Grove. Proceedings of the ASIS & T Annual Meeting 45. Silver Spring, MD: American Society for Information Science and Technology. doi:10.1002/meet.2008.1450450143

Saarikoski, Jyri, Jorma Laurikkala, Kalervo Järvelin, and Martti Juhola. 2011. "Self-organising Maps in Document Classification: A Comparison with Six Machine Learning Methods." In *Adaptive and Natural Computing Algorithms: 10th International Conference, ICANNGA 2011, Ljubljana, Slovenia, April 14-16, 2011; Proceedings,* ed. David Hutchison, Andrej Dobnikar, Uroš Lotrič, and Branko Šter. Lecture Notes in Computer Science 6593. Berlin: Springer: 260-9.

Salton, Gerard and Michael McGill. 1983. *Introduction to Modern Information Retrieval.* Auckland: McGraw-Hill.

Salton, Gerard. 1991. "Developments in Automatic Text Retrieval." *Science* 253: 974-9.

Sebastiani, Fabrizio. 2002. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys* 34, no. 1: 1-47.

Silvester, June P. 1997. "Computer Supported Indexing: A History and Evaluation of NASA's MAI System." In *Encyclopedia of Library and Information Science*, ed. Allen Kent. New York: Dekker, 61: 76-90.

Smiraglia, Richard P. and Xin Cai. 2017. "Tracking the Evolution of Clustering, Machine Learning, Automatic Indexing and Automatic Classification in Knowledge Organization." *Knowledge Organization* 44: 215-33.

Song, Wei, Jucheng Yang, Chenghua Li, and Sooncheol Park. 2011. "Intelligent Information Retrieval System Using Automatic Thesaurus Construction." *International Journal of General Systems* 40: 395-415.

Sousa, Renato Tarciso Barbosa de. 2014. "A representação da informação: classificação e indexação automática de documentos de arquivo [The Representation of Information: Automatic Classification and Indexing of Archives Records]." In *XV Encontro Nacional de Pesquisa em Ciência da Informação: além das nuvens, expandindo as fronteiras da Ciência da Informação, 27-31 de outubro em Belo Horizonte, MG,* ed. Isa M. Freire, Lilian M. A. R. Álvares, Renata M. A. Baracho, and Maurício B. Almeida. Belo Horizonte: ECI; UFMG, 798-811. http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt2

Souza, Renato Rocha and Koti S. Raghavan. 2014. "Extraction of Keywords from Texts: An Exploratory Study using Noun Phrases." *Informação & Tecnologia (ITEC)* 1, no. 1: 5-16.

Sparck Jones, Karen. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation* 28: 11-21.

Stanfill, Mary H., Margaret Williams, Susan H. Fenton, Robert A. Jenders, and William R. Hersh. 2010. "A Systematic Literature Review of Automated Clinical Coding and Classification Systems." *Journal of the American Medical Informatics Association* 17: 646-51.

Stevens, Mary E. 1965. *Automatic Indexing: A State of the Art Report.* National Bureau of Standards Monograph 91. Washington, D.C.: U.S. Government Printing Office.

Subramanian, Srividhya and Keith E. Shafer. 1998. "Clustering." https://web.archive.org/web/20040514080331/http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003409

Svarre, Tanja and Marianne Lykke. 2014. "The Role of Automated Categorization in E-government Information Retrieval." *Knowledge Organization* 41: 76-84.

Svenonius, E. 2000. *The Intellectual Foundations of Information Organization.* Cambridge, MA.: MIT Press.

Waltinger, Ulli, Alexander Mehler, Mathias Lösch, and Wolfram Horstmann. 2011. "Hierarchical Classification of OAI Metadata Using the DDC Taxonomy." In *Digital Libraries: Achievements, Challenges and Opportunities: 9th International Conference on Asian Digital Libraries, ICADL 2006, Kyoto, Japan, November 27-30, 2006; proceedings,* ed. Shigeo Sugimoto. Lecture Notes in Computer Science 6699. Berlin: Springer, 9-40.

Wan, Chin Heng, Lam Hong Lee, Rajprasad Rajkumar, and Dino Isa. 2012. "A Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-nearest Neighbor and Support Vector Machine." *Expert Systems with Applications* 39: 11880-8.

Wang, Jun. 2009. "An Extensive Study on Automated Dewey Decimal Classification." *Journal of the American Society for Information Science and Technology* 60: 2269-86.

Wartena, Christian and Maike Sommer. 2012. "Automatic Classification of Scientific Records Using the German Subject Heading Authority File (SWD)." In *SDA 2012: Semantic Digital Archives; Proceedings of the 2nd International Workshop on Semantic Digital Archives, Paphos, Cyprus, September 27, 2012;* ed. Annett Mitschick, Fernando Loizides, Livia Predoiu, Andreas Nürnberger. and Seamus Ross. http://ceur-ws.org/Vol-912/paper3.pdf

Weisser, Martin. 2015. *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis.* Hoboken, NJ: Wiley.

Willis, Craig and Robert M. Losee. 2013. "A Random Walk on an Ontology: Using Thesaurus Structure for Automatic Subject Indexing." *Journal of the American Society for Information Science and Technology* 64: 1330-44.

Witten, Ian H. *and* Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations.* San Diego, CA*: Academic Press.*

Yang, Yiming. 1999. "An Evaluation of Statistical Approaches to Text Categorization." *Journal of Information Retrieval* 1, nos. 1/2: 67-88.