Research on Automatic Classification of Documents in Library Environment: A Literature Review

Sanjay K. Desale* and Rajendra M. Kumbhar**

* Sanjay K. Desale, Jayakar Library, University of Pune, Pune, Maharashtra, India, <skdesale@unipune.ac.in>
** Rajendra M. Kumbhar, Department of Library and Information Science, University of Pune, Pune, Maharashtra, India, <rajendra_kumbhar@unipune.ac.in >

Sanjay Desale is assistant university librarian at Jayakar Library, University of Pune, India. At present, he is in charge of the library automation and circulation section of Pune University Library. Library automation in general and automated classification in particular are his research interests. He did his doctoral research on development of a semi-automated depth classification scheme for physics.

Rajendra Kumbhar is professor at the Department of Library & Information Science, University of Pune, India. Knowledge organization, in general, and thesaurus and classification, in particular, are his teaching and research interests. He did his doctoral research on development of a depth classification scheme and thesaurus of library and information science. Recently, he has published a book entitled *Library Classification Trends in the 21st Century*.

Desale, Sanjay K., and Kumbhar, Rajendra M. **Research on Automatic Classification of Documents in Library Environment: A Literature Review.** *Knowledge Organization. 40(5)*, 295-304. 78 references.

ABSTRACT: This paper aims to provide an overview of automatic classification research, which focuses on issues related to the automatic classification of documents in a library environment. The review covers literature published in mainstream library and information science studies. The review was done on literature published in both academic and professional LIS journals and other documents. This review reveals that basically three types of research are being done on automatic classification: 1) hierarchical classification using different library classification schemes, 2) text categorization and document categorization using different type of classifiers with or without using training documents, and 3) automatic bibliographic classification. Predominantly this research is directed towards solving problems of organization of digital documents in an online environment. However, very little research is devoted towards solving the problems of arrangement of physical documents.

Received 26 April 2013; Revised 12 July 2013; Accepted 31 July 2013

1.0 Introduction

Document classification is one of the prime functions of all libraries and information centers. For performing this activity, library professionals have designed and developed classification schemes. Melvil Dewey, Charles Ami Cutter, H. E. Bliss, S. R. Ranganathan, and others have contributed immensely to the world of bibliographic classification. Human beings have a strong attraction for developing and using automatic systems for their activities in various walks of life. Document classification is not an exception to this. Many library professionals believe that automatic classification will help in classifying more effectively, quickly, and accurately. Due to the information explosion, the printed manual classification schemes are becoming bulky and thereby expensive and unmanageable. Library professionals have invested their time in designing automatic document classification schemes as they help in standardizing the classification procedure. Standardization of classification helps in constructing uniform class numbers, which further helps in locating pinpointed information and documents.

Library professionals have been studying automatic document classification systems for quite some time. During their research, they have developed various types of classification schemes by applying various methods and techniques. They have applied newly designed systems with varying degree of success. Yet we are far away from having full-fledged automatic document classification schemes. During doctoral research on this topic, it was noticed that no literature review has been written so far covering the literature on automatic document classification. This article reviews this literature. The main purpose of this review is to identify trends in automatic document classification. It also aims to identify unattended facets by researchers in this area.

2.0 Methodology

This is a review of literature on automatic classification with particular focus on classification in a library environment. As a first step, the literature on the topic was searched through journal articles, books, theses, and other forms of documents. Only literature published in the English language was considered for this purpose. Automatic document classification has been a topic of interest for library professionals since the 20th century. In order to cover the maximum amount of literature, no time limit was applied in covering literature in this review. Obviously, most of the literature on this topic has been published in the late 20th century and the beginning of the 21st century. Literature published in printed form as well as in electronic form is covered by this review. Like most other subjects, much of the library and information science literature is available online. Keeping this fact in mind and to ensure optimum coverage various online databases were searched to access literature on automatic document classification. LISA and LISTA are the prime databases covering most literature on library and information science. Therefore literature was searched in these two databases. Emerald Insight publishes a large number of journals in library and information science, so this database was scanned thoroughly. Science Direct, Taylor and Francis, and Sage Publications databases were also deliberately searched as these databases also cover some of the reputable library and information science journals. A large amount of literature is now published in open access journals, which are extensively covered by Google Scholar; therefore this database was used to search literature on the topic of this review. JSTOR database was also searched so as to ensure coverage of retrospective literature on this topic.

Online databases were used extensively to search and retrieve the required literature. In order to ensure the retrieval of relevant literature, it was essential to use appropriate keywords. For this purpose, first an extensive list of keywords was prepared. This list included: classification, library classification, bibliographic classification, document classification, automatic and automated classification, automation and classification, special and general classification, depth classification, enumerative and faceted classification. In addition to these, modern terminology such as knowledge organization, thesauri, taxonomy, ontology, semantics, computerized classification, natural language processing, artificial intelligence, and text categorization were used. These keywords were used to search literature from the aforementioned databases by using various Boolean combinations.

Most of the documents were collected in full-text format. These were studied and analyzed so as to know the nature of research, methodology adopted, and the conclusions arrived at. The limitations of the concerned research were also taken into consideration while studying the given research. A review based on the minute study of the relevant document is presented in the following paragraphs. The review is presented in a thematic manner. Seven broad themes are identified for presenting the review. Due to the nature of the subject, some overlap over the themes covered is inevitable.

3.0 Need of automated classification

According to Pong et al. (2008, 213):

> With the explosive growth in the number of electronic documents available on the Internet and digital libraries, it is increasingly difficult for library practitioners to categorize both electronic documents and traditional library materials using just a manual approach .... To improve the effectiveness and efficiency of document categorization at the library setting, more in-depth studies of using automatic document classification methods to categorize library items are required.

Mengle and Goharian (2009, 1037) also wrote that "with the increasing number of digital documents the ability to automatically classify those documents both efficiently and accurately is becoming more critical and difficult." Wang (2009, 2269) suggested that "the automation of the subject-classification process has long been needed to increase the efficiency of catalog production and to free classifiers from the heavy workload, especially in today's information explosion age." On the information explosion, Wiggins (2005) reported that the Library of Congress (LC) collection exceeded 130 million items and is increasing at a speed of more than two million items per year. The OCLC Online Computer Library Center (2010) research group stressed that, "though cataloging and classification requires

https://doi.org/10.5771/0943-7444-2013-5-295 https://www.inlibra.com/de/agb - Open Access -

expert intellectual effort, we recognize that at least some of the work must be automated if we hope to keep pace with cultural change." Berners-Lee (2001) argued that the current web would be transformed into a more intelligent semantic web when it is augmented with data for automated processing. Ranganathan (1965a, 541) suggested that the "future work of FID/CR should be to encourage the design to improved schemes of classification, whether they are going to be faceted, analytic synthetic, or whether new brand it will be. In doing this, there should be frequent consultations with the machine specialists." The need for automated classification was realized by library and information science professionals and many efforts have been made from since the early 1970s.

4.0 Early efforts in automated classification

"Early efforts in automatic subject classification (including subject indexing) date back to the 1970s" (Wang 2009, 2269), and various approaches have been explored, including rule based methods, statistics based methods and information retrieval based methods. In rule based methods, the first study reported in the literature is Schiminovich (1971) based on a pattern discovery algorithm, which used citation content of the document and bibliographic links among papers. However, the study was confined to the retrieval of the document; Schiminovich claimed a 100% recall and relevance ratio. Cahn and Herr (1978, 11) declared "automatic document classification to be an enlightening and achievable goal within artificial intelligence." Gopinath and Prasad (1994) presented the model for knowledge representation for analytic synthetic classification. In statistics based methods, Field (1975) published an article based on statistical ranking and weighing methods for automatically assigning subject headings. Cheng and Wu (1995, 289) introduced automated classification system for school libraries, claiming "80% correctness in automatic classification and a cost reduction of 75% compared to manual classification." In retrieval based methods, Sudha (1986) emphasized the use of artificial intelligence in information retrieval systems. Larson (1992, 147) reported that "the most effective methods combined the use of the first subject heading as the representative for the item to be classified, with either the complete [*Library of Congress Subject Headings* (*LCSH*) phrase] or keywords from the heading with plural suffixes converted to singular form, and the use of one of the probabilistic search methods." Vizinne-Goetz (1996) highlighted the use of classification schemes for the retrieval of online information. Carpineto and Romano (1994) highlight the deficiencies of traditional online documentation systems and also gave architecture of intelligence help system. Almost all of these early efforts were mainly directed towards term classification—grouping and arraying terms for machine readable databases to improve computer search efficiency or to retrieve information from the internet. However more concentrated efforts are seen in recent literature.

5.0 Recent efforts in automated classification

More recently, considerable research has been undertaken, but this work is dominated by text categorization (Sebastiani 2002) and document categorization (Kim and Choi 2007) in an electronic environment. However very little research has been done for atomic classification of physical documents in a library environment. Almost all classification schemes have been tried for designing automated classification schemes, GERHARD project used the Universal Decimal Classification (UDC) (Toth 2002), while the SCORPION (OCLC 2012a) project employed the *Dewey Decimal Classification* (*DDC*). The Library of Congress *Classification* (*LCC*) has also been used for experimental document classification (Frank and Paynter 2004). But "at present there do not appear to be any practical examples where library classification systems have been completely overtaken by automatic methods. However there is an increasing interest in developing such systems" (Toth 2002, 48).

5.1 Text categorization:

Automatic classification research has made prominent progress in the text-categorization (TC) field in recent years (Sebastiani 2002), especially the supervised machine learning approach, which has shed new light on the resolution of this problem. TC, also called text classification or text spotting, is the activity of labelling natural-language texts with thematic categories from a predefined set. "The most frequent approach to automated classification is machine learning. It, however, requires training documents and performs well on new documents only if these are similar enough to the former" (Golub et al. 2007, 248). Automatic text classification "systems require extensive information about the book in machine-readable form, for example, an abstract, table of content, or the complete text of the work. Providing such information when it is not available beforehand is costly and impractical" (Avila-Arguelles et al. 2010). Aiolli et al. (2009, 578) "addressed the problem of how to learn a classifier that distinguishes between the primary and the secondary categories of a document, and argued that this task deserves to be explicitly tackled by TC research." Mengle and Goharian (2010) wrote that "effectively discovering relationships among categories is useful in the field of text mining and text catagorization." Unlike the use of a concept or category hierarchy, Mengle and Goharian represent relationships among categories using a graph structure called a relationship net. Liu (2010, 308) argues that "the semantics of each term in a document is essential for the classification of the document, while the term's semantics in a document often heavily depends on its context (neighboring) terms." Kim and Choi (2007, 1200) "propose a patent document catagorisation method that uses the k-NN (k-Nearest Neighbour) approach."

Most of current research in automated classification is in text categorization. Major techniques used in text categorization are: support vactor machine models, k-nearest neighbour, machine learning, frequency measure, and weighing technique. Different classifiers like probabilistic classifiers, decision tree classifiers, naïve Bayes classifiers, and linear classifiers are used for text categorization. The major problem with these techniques is that they require data in machine readable form, training documents, and still they often fail to produce unique class numbers needed to place a document on library shelves.

5.2 Document classification

The best approach for document classification without training documents, which could be useful for document classification in a library environment, was suggested by Golub et al. (2007). According to Golub et al. (2007, 248), "in document classification, matching is conducted between a controlled vocabulary and text of documents to be classified. A major advantage of this approach is that it does not require training documents. If using a well-developed classification scheme, it will also be suitable for subject browsing in information retrieval systems." Slavic (2007, 581) suggests that "classifications have the advantage of supporting systematic organization." In the past, we also had knowledge organization systems called thesauro-facets (Aitchison et al. 1969) or classaurus (Bhattacharyya 1982). Slavic also wrote (2007, 582) that "modern analytico-synthetic and faceted classifications have greater potential in knowledge organization." More recently, faceted classification has been used in subject directories and search engines (Ellis and Vasconcelos 2000). Kovacevic et al. (2011) proposed a system for the fully automated extraction of metadata from scientific publications. Unlike the use of full-text in text categorization, metadata can be used for automated document classification. Metadata are available in online public access catalogues of many libraries. Classification schemes can be used in place of training documents to produce class numbers.

5.3 Hierarchical classification schemes and automation:

It is accepted fact in the literature of library and information science that traditional classification schemes are also useful for automated classification (Kim and Lee 2002; Pong et al. 2008; Wang 2009). Van der Walt (1997) highlighted the advantages of library classification schemes for organization of information resources in the web environment. According to Van der Walt, the knowledge organization tools developed and used by web search engines often feature shallow hierarchies and uneven coverage of topics. On the other hand, web search engines often respond to popular topics more quickly than traditional library knowledge organization tools do. In the context of hierarchical browsing based on a classification scheme, having too many classes assigned to a document would place one document in many different places, which would create the opposite effect of the original purpose of a classification scheme (grouping similar documents together) (Golub et al. 2007). In this context, the use of a classification scheme is mostly confined to the hierarchical arrangement of digital documents in the network environment only.

Apart from aforementioned problems of hierarchical browsing, there are some other problems with library classification schemes that are discussed in the literature. Frank and Paynter (2004) address the problem of automatically assigning LC*C* to a work given its set of *LCSH* headings. LC*C*s are organized in a tree. The root node of this hierarchy comprises all possible topics, and leaf nodes correspond to the most specialized topic area defined; it is difficult to automatically identify leaf nodes corresponding to root nodes. Wang (2009) also argues that the number synthesis process in *DDC* is not reversible, and it is hard, even for a professional, to identify the boundary between the base number and facet notation because of the inconsistent usage of facet indicators; sometimes facet indicators are built into the base numbers, sometimes the ending zeros of a base number are dropped off (e.g., 500=0785=507.85), sometimes 0 is used as facet indicator (for general subdivisions), and sometimes 2 is used. Slavic (2007) also has the view that the disciplinary structure of decimal classifications such as *DDC* and UDC with ten main classes is very poorly equipped to properly represent the universe of knowledge.

However Golub et al. (2007, 262) advocate "classifying documents into classes of well-developed classification schemes … suitable for subject browsing, unlike automatically developed controlled vocabularies or home-grown directories often used in document clustering and text categorization." Chan (2001) also wrote: "when subject categorization devices first became popular among Web information providers, they resembled broad classification schemes, but many were lacking the rigorous hierarchical structure and careful conceptual organization found in established schemes."

Only a few automated document classification systems in the literature are constructed based on standard document classification schemes. German Harwest Automated

https://doi.org/10.5771/0943-7444-2013-5-295 https://www.inlibra.com/de/agb - Open Access -

Retrieval and directory created a HARVEST based robot-generated index of Web resources in Germany using indexing and automatic classification and provided a search and hierarchical browsing facility based on the UDC, while the Scorpion project employed the *DDC* (Toth 2002). The Scorpion Open Source project offers software that implements a system for automatically classifying Web-accessible text documents. "Scorpion is intended for use by investigators who have a machine-readable subject classification scheme or thesaurus and wish to incorporate it into an automatic classification system" (OCLC 2012). "An experimental automatic document classification system was also built using the L*CC* scheme .... The procedure uses machine learning techniques and training data from a large library catalog to learn a classification model mapping from sets of *LCSH* to nodes in the L*CC* tree" (Frank and Paynter 2004, 214). Frank and Paynter (2004) showed accuracy on an independent collection of 50,000 *LCSH*/L*CC* pairs. OCLC has developed a tool, Classify, a FRBR-based research prototype for applying classification numbers, which provides a summary of all the class numbers applied to the work (Vizine-Goetz 2010). In Classify, one can also view a series of charts that show the top ten assigned classes from the *DDC*, the L*CC*, and the National Library of Medicine (NLM) Classification.

"However, none of these document classification systems was constructed based on machine learning techniques." (Pong et al. 2008, 215). Soergal et al. (2004) also pointed out that "existing classification schemes and thesauri lack well-defined semantics and structural consistency." Pong et al. (2008, 214) note that "in the literature of library and information science, the need to combine electronic documents with traditional library materials has inspired continuous discussions on the refinement of existing manual classification schemes." Hunter (2009) writes that the principles upon which Colon Classification is based are important. It is clear from literature that hierarchical library classification schemes are useful for hierarchical browsing but are of little use in automatically producing class numbers. Faceted classification based on sound principles might be useful in automation.

5.4 Facet classification and automation

Broughton (2005) suggested that a faceted system is more suitable in electronic environments than enumerative and pre-coordinated systems for information retrieval. Kim and Lee (2002) designed a knowledge base for an automatic classification in the library and information science field, by using the facet classification principles of Colon Classification. Broughton (2008) thinks that facet analysis provides a sound basis for structuring a variety of knowledge organization tools. Uddin and Janecek (2007, 231) found that "faceted classification allows the users of a website to access information more efficiently than the simple taxonomic hierarchy of information object." Devadason et al. (2002, 66) "attempted to describe an experimental system designed to organise and provide access to web documents using a faceted pre-coodinate indexing system based on a deep structure indexing system derived from POPSI (Postulate Based Permuted Subject Indexing) of Bhattacharya, and the facet analysis and chain indexing system of Ranganathan." Panigrahi and Prasad (2007) demonstrated the techniques of fixing the facet sequence in developing an automatic classification system to construct classification numbers for document titles, which appear in natural language. Hunter, (2009, 8) quoting Clifton, wrote that the "nature of faceted classification enables it to be more easily interpretable by both human beings and computers." These studies suggest that little success is achieved in automatic class number generation by using faceted classification schemes. Rather, a faceted structure is more suitable for automated classification; it also needs a relatively small vocabulary for knowledge representation.

5.5 Use of natural language and artificial intelligence in automatic classification

Panigrahi (2000) argued that natural language processing could be used in the automatic identification of noun phrases from the expressive title. Kim and Lee (2002) argued that book titles usually have an immediate connection to their contents in that they often encapsulate the entire work. Wang (2003) also has similar views that the title of a document usually summarizes its contents and reveals its central topics. Kwok (1975) also thinks that, since the introduction of KWIC indexes, there are reasons to believe that authors are writing more descriptive and meaningful titles than before, and this gives added confidence in their use. For Dutta et al. (2008), keywords indicate core concepts and central fields of concern. The keywords are building blocks of the 'descriptors' or 'subject headings' because subject headings are composed of several keywords. Studies on indexing show significant variation in the use of keywords selected by different indexers to represent the same topic or document (Bertrand and Cellier 1995). However, a widely accepted belief in text categorisation studies is that using individual words to represent a document is better than using n-gram phrases (Lewis 1992). Avila-Arguelles et al. (2010) presented an experiment on supervised classification of books by using their titles only, which would allow massive atomic indexing. They proposed a new text comparison measure, which mixes two well-known text classification techniques: the Lesk voting scheme and the term frequency (TF). In addition, they experimented with different weighting as well as logical combinatorial methods such as ALVOT in order to determine the contribution of the title to the correct classification, which they found was approximately one third.

5.6 Interactive classification

Interactive classification has been preferred for automation of classification by some of the experts; prominent among them is Wang (2009), who demonstrated the advantage of interactive classification. Sparck Jones (2005) also advocated a posteriori classification of documents. Within automatic classification, an interactive classification has produced better results than non-interactive automatic classifications.

6.0 Need of methodology for automated classification

Library and information science literature lacks an established methodology for designing automated book classification schemes. Hjørland and Pedersen (2006) argued that further progress in IR was impeded by a lack of a substantive classification theory. Karamuftuoglu (2007, 1978) also stressed the need for substantive classification theory by saying "document classification should be based on (informed by) the theories/paradigms that exist in a given domain or discipline, as well as tasks and goals of specific user groups." Most of the library classification schemes were developed before the invention of the computer; therefore, a methodology for automated classification was not developed.

7.0 Types of library classification schemes

According to Slavic (2007, 585) "there are three types of knowledge organization structures that are relevant in knowledge mediation: taxonomic, aspect (i.e. disciplinary-based), and phenomena-based." Classificatory structures are like, "hierarchies, trees, paradigms, and facet analysis" (Kwasnik 1999, 22). There are many classification schemes and "most of the widely used documentary classifications are disciplinary, i.e. aspect classifications" (Slavic 2007, 586). Many early classification schemes were developed on an adhoc basis has weak theoretical principles (Broughton 2004), whereas the more frequently implemented *DDC* has greater theoretical rigour (Giess et al. 2008). UDC was also based on the *DDC*'s original structure and notations, but differs from enumerative Dewey by virtue of its synthetic nature. Ranganathan's Colon Classification (Ranganathan 1965b) and the second edition of Bliss Classification (BC2) (Broughton 2001) are both faceted classifications. The ideas of Colon Classification were revisited by the Classification Research Group in the 1960s, which led to generation of BC2 in the late 1970s (Giess et al. 2008).

8.0 Faceted classification

"A perfect solution of all problems connected with the storing and retrieving of information in chemistry can be expected only from a consistantly analytico-synthetic classification, the fundamentals of which are laid down in Ranganathan's *Prolegomena to library classification*" (Fugmann 1965, 1). Ranganathan (1960) introduced an idea of dividing and organizing complex subjects by facets in the 1930s. He describes five fundamental facets, called PMEST in his Colon Classification. The idea was well accepted in literature by the international community. Grolier (1965) quoted Gardin's grammatical categories, which are not of the same nature as Ranganathan's categories, but which serve as orientation for relations. Philosophical interest in categories may be traced back to Aristotle, who in his treatise Categories (1963 ed.), attempts to enumerate the most general kinds as: substance, quantity, relation, place, date, posture, state, action, and passion. The theory of faceted classification by the Classification Research Group (Vickery 1960), and used in BC2 published in 1970, extends Ranganathan's original five categories to thirteen categories (Thing – kind – part – property – material – process – operation – patient – product – by-product – agent – space – time). However, Hjørland (2008) notes that "Vickery's expansion of the number of categories may imply that there is not a fixed set of categories in the world."

In the Dorking conference, it was generally agreed that future classification schemes should preferably be of a faceted kind (Neelameghan 1965). The resilience of a freely faceted classification is greater than that of an almost faceted classification or a rigidly faceted classification (Ranganathan 1989). Gnoli and Mei (2006) suggest joining the merits of free classification with those of faceted classification to form a freely faceted classification. The term "Free Classification" was first introduced by Gardin (1965) in the article which gives the respective merits of Free Classification and Faceted Classification with special regard to their implementation on computers.

9.0 Depth classification

For many years, it has been argued by many experts that none of the general classification schemes are satisfactory. Foskett (1964) quoted E. J. Coate, who drew attention to several of the problems arising out of the inability of the well-known general schemes to cater to the complexities of modern knowledge and the demands of modern library services. Ranganathan (1965c) argued that the impracticability of general classification schemes drives one to the expediency of building a separate special classification for each different small subject area in which new formations make the respective regions of general schemes appear markedly unhelpful. Mills (1964) explained the detailed inadequacies of general classification schemes, arguing that the general classification schemes have assimilated new knowledge only after the fact and frequently in a manner which users could not have foreseen. This failure is partly due to wrong priorities which have allowed brevity, simplicity, and expressiveness of notation to prevail over classification ordinal structure, and partly due to arbitrariness in the initial scheme (Coates 1964).

Ranganathan believed that depth classification schemes of micro subjects function as a link in the chain of communication needed to prevent the reversion of relay-research into research by isolated individuals (Ranganathan 1989). Ranganathan defined depth classification as "a scheme of classification fitted to reach co-extensiveness and expressiveness in the classification of micro thoughts having many rounds and levels of facets and isolates of high orders in any or all of them" (Ranganathan 1957, 241). Vickery (1960) published a manual for the construction of special schemes. Kumbhar (2005) narrated the experience of constructing a thesaurus of LIS terms by using speciator-based faceted depth classification schedules. He also explained the advantages of this method in establishing various thesaural relationships. Gopinath (1986) published a manual which gives instructions for the construction of depth classification schedules. Kumbhar (2002, 28) emphasized the need for depth classification in his thesis by arguing that:

> Any general, broad classification scheme cannot truly serve both the self- arrangement and documentation. The depth classification schemes fulfill both these requirements, for it lists micro subjects and provides rules for synthesizing the various terms representing the subject. Further a term in a compound phrase may form part of many combinations and it is only the depth classification's faceted nature, which provides complete flexibility in coordination of terms.

The first round of routine design of depth classification for diverse subject fields was thought to require about 5,000 man-years of work (Ranganathan 1964). Gopinath (1975) suggested, however, that there were several initial steps which are common for construction of both a thesaurus and a classification scheme.

Recently Pong et al. (2008) argued the *LCC* classes can be broken down into deeper levels, so that more specific areas of the subject can be covered. Panigrahi and Prasad (2007, 42) concluded by saying that "Ranganathan's idea of defining isolates in the line of meccano set is the best suited and highly computer compatible to develop automated classification system using artificial intelligence techniques." However, methodology for designing such classification schemes has not been developed (Golub et al. 2007). Ranganathan's canons, postulates and principles and Spiteri's later revisions define the requisite properties that a faceted structure should have, but do not provide a methodology that can be followed in order to arrive at this structure (Giess et al. 2008). Neelameghan and Gopinath (1965) suggested a pragmatic approach for designing a depth classification scheme. Beghtol (1986) summarized "literary warrant," an empirical principle widely recognized in knowledge organization, first coined by Hulme in his paper "Principles of Book Classification." According to Hulme, the definition of a class heading should rest upon a purely literary warrant. A class heading is warranted only when literature has been shown to exist in book form, and the test of the validity of a heading is the degree of accuracy with which it describes the area of subject matter common to the class. Beghtol (1986) therefore suggests that the definition of literary warrant "may be described as the plotting of areas pre-existing in the literature."

10.0 Conclusion

This review indicates that basically three types of research are ongoing on automatic classification: 1) hierarchical classification by using different library classification schemes, 2) text categorization and document categorization by using different type of classifiers with or without using training documents, and 3) automatic bibliographic classification. Predominantly, this research is directed towards solving problems of organization of digital documents in an online environment. However, very little research is devoted towards solving the problems of arrangement of physical documents in a library environment. This review also suggests that an automated classification scheme can be designed by using natural language and artificial intelligence.

Also highlighted is the fact that Ranganathan's canons, postulates, and principles define the requisite properties that a faceted structure could have, but do not provide a methodology that can be followed in order to arrive at these structures. Faceted classification schemes are more suitable for automatic classification than enumerative classification schemes. It is also established from the literature that, because of the exponential growth of literature, there is need for depth classification schemes. Very few studies are found in the published literature, which suggests use of depth classification and thesauri for the development of automated classification of physical documents in libraries.

This review also suggests that the amount of research on text categorization is increasing at a noticeable pace. Compared with this not much research is carried out on

https://doi.org/10.5771/0943-7444-2013-5-295 https://www.inlibra.com/de/agb - Open Access -

automatic document classification. It is worth studying whether this is due to the simplicity of text categorization and the complexity of document classification, or due to some other reason. The present available research also shows that most of the research on text categorization is carried out without using principles of document classification. Research based on the collaboration of library and information science professionals and information technology experts will probably bring more success in automatic document classification.

References

Aiolli, Fabio, Cardin, Riccardo, Sebastiani, Fabrizio and Sperduti, Alessandro. 2009. Preferential text classification: learning algorithms and evaluation measure. *Information retrieval* 12: 559-80.

Aristotle. 1963. *Categories*. Oxford: Clarendon Press.

Aitchison, Jean, Gomersall, Alan and Ireland, Ralph. 1969. *Thesaurofacet: a thesaurus and faceted classification for engineering and related subjects*. England: English Electric Company Ltd.

Avila-Arguelles, Ricardo, Calvo, Hiram, Gelbukh, Alexander and Godoy-Calderon, Salvador. 2010. Assigning Library of Congress Classification codes to books based only on their titles. *Informatica* 34: 77-84.

Beghtol, Clare. 1986. Semantic validity: concepts of warrant in bibliographic classification systems. *Library resources & technical services* 30: 109-25.

Berners-Lee, Tim. 2001. The semantic web - computers navigating tomorrow's web will understand more of what's going on - making it more likely that you will get what you really want. *Scientific American* 284: 34-43.

Bertrand, Annick and Celler, Jean-Marie. 1995. Psychological approach to indexing effects of the operators expertise upon indexing behaviour. *Journal of information science* 21: 459-72.

Bhattacharyya, G. 1982. Classaurus: its fundamentals, design and use. In Dahlberg, Ingetraut, ed., *Universal classification, subject analysis and ordering systems: proceedings of the 4th international study conference on classification research, Augusburg*. Frankfurt: Indeks Verlaag, pp. 139-48.

Broughton, Vanda. 2001. Faceted classification as a basis for knowledge organisation in a digital environment; the Bliss Bibliographic Classification as model for vocabulary management and the creation of multi-dimensional knowledge structures. *New review of hypermedia and multimedia* 7: 67-102.

Broughton, Vanda. 2004. *Essential classification*. London: Facet Publishing.

Broughton, Vanda. 2005. The need for faceted classification as a basis of all methods of information retrieval. *Aslib proceedings: new information perspectives* 58: 49-72.

Broughton, Vanda. 2008. A faceted classification as the basis of a faceted terminology: conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification, 2nd Edition. *Axiomathes* 18: 193-210.

Cahn, D.F. and Herr, J. 1978. Automatic document classification based on expert human decisions. In Brenner, E.H. and Plains, W., eds., *Proceedings of the ASIS annual meeting 1978* White Plains, NY: Knowledge Industry Publications, pp. 63-66.

Carpineto, Claudio and Romano, Giovanni. 1994. Dynamically bounding browsable retrieval spaces: an application to Galois Lattices. In *Proceedings of RIAO 94: intelligent multimedia information retrieval systems and management New York*, pp. 520-33.

Chan, Lois Mai. 2001. Exploiting LCSH, LCC, and DDC to retrieve networked resources: issues and challenges. In *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium*. Washington, DC: Library of Congress, pp. 159-78. Available http://www.loc.gov/catdir/bibcontrol/chan_paper.html accessed 21 August 2013.

Cheng, Patrick T.K. and Wu, Albert K.W. 1995. ACS: an automatic classification system. *Journal of information science* 21: 289-99.

Coates, E.J. 1964. *CRG proposals for a new general classification*. Hertfordshire: The Library Association.

Devadason, Francis J., Intaraksa, Neelawat, Patamawongjariya, Ponprapa and Desai, Kavita. 2002. Faceted indexing based system for organising and accessing Internet resources. *Knowledge organization* 29: 65-74.

Dutta, Bidyarthi, Mujumdar, Krishnapada and Sen, B. K. 2008. Classification of keywords extracted from research articles published in science journals. *Annals of library and information studies* 55: 317-33.

Ellis, David and Ana, Vasconcelos. 2000. The relevance of facet analysis for world wide web subject organisation and searching. *Journal of internet cataloguing* 2: 97-114.

Field, B.J. 1975. Towards automatic indexing: automatic assignment of controlled language indexing and classification from free indexing. *Journal of documentation* 31: 246-65.

Foskett, D. J. 1964. *Origins of conference*. Hertfordshire: The Library Association.

Frank, Eibe and Paynter, Gordon W. 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.

Fugmann, R. 1965. Experiences with a faceted classification in organic chemistry using computers. In Atherton, P. ed., *Classification research: proceeding of the second international study conference*. Copenhagen: Munksgaard, pp. 341-67.

Gardin, J. C. 1965. Free classifications and faceted classifications. In Atherton, P., ed., *Classification research: proceeding of the second international study conference*. Copenhagen: Munksgaard, pp. 161-8.

Giess, M. D., Wild, P. J. and Mcmahon, C. A. 2008. The generation of faceted classification schemes for use in the organisation of engineering design documents. *International journal of information management* 28: 379-90.

Gnoli, Claudio and Mei, Hong. 2006. Freely faceted classification for web-based information retrieval. *New review of hypermedia and multimedia* 22: 63-81.

Golub, Koraljka, Hamon, Thierry and Ardo, Anders. 2007. Automated classification of textual documents based on a controlled vocabulary in engineering. *Knowledge Organization* 34: 247-63.

Gopinath, M. A. 1975. Thesaurus and classification scheme: a study of the compatibility of the principles for consturction of thesaurus and classification scheme. In *Thesaurus in Information Systems*. Bangalore: DRTC, pp. A37-A50.

Gopinath, M. A. 1986. Construction of depth version of colon classification: a manual. Michigan: Wiley Eastern.

Gopinath, M.A. and Prasad, A.R.D. 1994. A knowledge representation model for analytico- synthetic classification. In Albrechtsen, Hanne and Oernager, Susanne, eds., *Knowledge organization and quality management: proceedings of the third international ISKO conference 20-24 June 1994 Copenhangen, Denmark*. Frankfurt/Main: Index Verlag, pp. 320-7.

Grolier, E. D. 1965. Current trends in theory and practice of classification. In Atherton, P., ed., *Classification research: proceeding of the second international study conference*. Copenhagen: Munksgaard, pp. 9-14.

Hjørland, Birger and Pedersen, Karsten Nissen. 2006. A substantive theory of classification for information retrieval. *Journal of documentation* 61: pp. 582-97.

Hjorland, Birger. 2008. Facet, facet analysis and the facet-analytic paradigm in knowledge organisation (KO). Available http://www.iva.dk/bh/lifeboat_ko/CONCEPTS/facet_and_facet_analysis.htm.

Hunter, Eric J. 2009. *Classification made simple: an indroduction to knowledge organisation and information retrieval*. Surray: Ashgate.

Karamuftuoglu, M. 2007. Need for a systemic theory of classification in information science. *Journal of the American Society for Information Science and Technology* 58: 1977-87.

Kim, Jeong-Hyen and Lee, Kyung-Ho. 2002. Designing a knowledge base for automatic book classification. *Electronic library* 20: 488-95.

Kim, Jae-Ho and Choi, Key-Sun. 2007. Patent document categorization based on semantic structural information. *Information processing and management* 43: 1200-15.

Kovacevic, Aleksandar, Ivanovic, Dragan, Milosavljevi, Branko, Konjovic, Zora and Surla, Dusan. 2011. Automatic extraction of metadata from scientific publications for CRIS systems. *Program: electronic library and information systems* 45: 376-96.

Kumbhar, R. M. 2002. *Construction of a vocabulary control tool (thesaurus) for library and information science*. Ph.D. dissertation. Aurangabad: Dr.Babasaheb Ambedkar Marathwada University.

Kumbhar, Rajendra. 2005. Speciator based faceted depth classifications applications in thesaurus construction. *Annals of library and Information Studies* 52: 15-24.

Kwasnik, Barbara H. 1999. The role of classification in knowledge representation and discovery. *Library trends* 48 no.1: 22-47.

Kwok, K. L. 1975. The use of title and cited titles as document representation for automatic classification. *Information processing and management* 11: 201-6.

Larson, Ray. 1992. Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science* 43: 130-48.

Lewis, David D. 1992. An evaluation of phrasal and clustered representations on a text categorization task. In Belkin, N., ed., *Proceeding of the 15th international ACM SIGIR conference on research and development in information retrieval*. New York: ACM Press, pp. 37-50.

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.

Mengle, Saket S.R. and Goharian, Nazli. 2009. Ambiguity measure feature-selection algorithm. *Journal of the American Soiciety for Information Science and Technology* 60: 1037-50.

Mengle, Saket S.R. and Goharian, Nazli. 2010. Detecting relationships among categories using text classification. *Journal of the American Society for Information Science and Technology* 61: 1046-61.

Mills, Jack. 1964. *Inadequacies of existing general classification schemes*. Hertfordshire: The Library Association.

Neelmeghan, A. 1965. New developments in library classification in India. In Atherton, P. ed., *Classification research: proceeding of the second international study conference*. Copenhagen: Munksgaard, pp. 503-23.

Neelameghan, A. and Gopinath, M. A. 1965. Pragmatic approach in the design of a depth classification schedule: a case study. *Library Science* 2: 55-68.

Online Computer Library Center [OCLC]. 2010. Automatic Classification Research. Available http://www.oclc.org/reserach/activities/past/orprojects/auto_class/defalult.htm.

Online Computer Library Center [OCLC]. 2012. Scorpion. Available http://www.oclc.org/research/activities/past/orprojects/scorpion/default.htm.

Panigrahi, Pijushkanti and Prasad, A. R. D. 2007. Facet sequence in analytico synthetic scheme: a study for developing an AI based automatic classification system. *Annals of library and information studies* 54: 37-43.

Panigrahi, P. K. 2000. An artificial intelligence approach towards automatic classification (part 2). *IAALIC bulletin* 45: 104-18.

Pong, Joanna Yi-Hang, Kwok, Ron Chi-Wai, Lau, Raymond Yiu-Keung and Wong, Percy Ching -Chi. 2008. A comparative study of two automatic document classification methods in a library settings. *Journal of information science* 34: 213-30.

Ranganathan, S. R. 1957. *Prolegomena to library classification.* 2nd ed. London: The Library Association.

Ranganathan, S. R. 1960. *Colon Classification, basic classification.* 6th ed. New York: Asia Publishing House.

Ranganathan, S. R. 1964. Design of depth classification: methodology. *LibraryScience* 1: 39.

Ranganathan, S. R. 1965a. Discussion on Neelmeghan and Rigby. In Atherton, P., ed., *Classification research: proceeding of the second international study conference.* Copenhagen: Munksgaard, pp. 540-2.

Ranganathan, S. R. 1965b. Library classification through a century. In Atherton, P., ed., *Classification research: proceeding of the second international study conference.* Copenhagen: Munksgaard, pp. 15-35.

Ranganathan, S. R. 1965c. General and special classifications. In Atherton, P., ed., *Classification research: proceeding of the second international study conference.* Copenhagen: Munksgaard, pp. 81-93.

Ranganathan, S. R. 1989. *Prolegomena to library classification.* 3rd ed. Bangalore: Sarada Ranganathan Endowment for Library Science.

Schiminovich, S. 1971. Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm. *Information storage and retrieval* 6: 417-35.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM computing surveys* 34: 1-47.

Slavic, Aida. 2007. On the nature and typology of documentary classifications and their use in a networked environment. *El profesional de la información* 16: 580-9.

Soergal, Dagobert, Lauser, Boris, Liang, Anita, Fisseha, Frehiwot, Keizer, Johannes and Katz, Stephen. 2004. Re-engineering thesauri for new applications: the AGROVAC example. *Journal of digital information* 4 no.4: 1-23.

Sparck Jones, Karen. 2005. Revisiting classification for retrieval. *Journal of documentation* 61: 598-601.

Sudha, S. 1986. Artificial intelligence in information retrieval system. *Library science with a slant to documentation* 23 no. 4: 214-22.

Toth, Eezsebet. 2002. Innovative solutions in automatic classification: a brief summary. *Libri* 52: 48-53.

Uddin, Mohammad Nasir and Janecek, Paul. 2007. Faceted classification in web information architecture: a framework for using semantic web tools. *Electronic library* 25: 219-33.

Van der Walt, Marthinus S. 1997. The role of information retrieval on Internet: some aspects of browsing lists in search engines. In *Knowledge organisation for information retrieval: proceedings of the 6th international study conference on the classification research.* Hague: FID, pp. 32-5.

Vickery, Brian C. 1960. *Faceted Classification: a guide to the construction and use of special schemes.* London: Aslib.

Vizinne-Goetz, Diane. 1996. Using library classification schemes for internet resources. Available http://staff.oclc.org/~vizine/Intercat/vizine-goetz.htm.

Vizine-Goetz, Diane. 2010. Classify: a FRBR-based research prototype for applying classification numbers. Available http://www.oclc.org/nextspace/014/research.htm .

Wang, June. 2003. A knowledge network constructed by integrating classification, thesaurus, and metadata in digital library. *The international information and library review* 35: 383-97.

Wang, June. 2009. An extensive study on automated Dewey Decimal Classification. *Journal of the American Society for Information Science and Technology* 60: 2269-86.

Wiggins, Beacher J. 2005. Annual report for acquisitions and bibliographic access directorate. Available http://www.loc.gov/catdir/bad05.pdf.