Automatic classification of office documents: Review of available methods and techniques

Savic, Dobrica

Automation is often considered to be a panacea for many administrative office tasks in today's world of increasing financial restrictions and difficulties. Classification of office documents, as one of the administrative functions carried out by every single organization or institution which sends and receives correspondence, is always labour intensive, time consuming and therefore expensive. This provides an important incentive for research efforts directed at designing a system to classify automatically the ever increasing amount of officially generated in-coming and out-going mail. A well founded and properly designed automatic system for classification of office documents can carry out the classification reliably, saving time and effort, while cutting down the cost which accompanies today's manual classification systems. Academic environments have demonstrated theoretical, while business has demonstrated pragmatic, interests for new developments and applications of modern information technology and methodology in this area.

This study examines the present status of available knowledge, and the state of the art of the methodology used for automatic classification of office documents. There is a wide spectrum of currently available and potential techniques. Their applicability to the task of automatic classification is examined through a literature review. In addition to reviewing the classic methods and techniques, the main focus of this research is on the application of artificial intelligence toward the development of an efficient and effective automatic classification system for office documents.

Classification of office documents is performed by almost all organizations and institutions. It is regarded as an act of identifying documents or records in accordance with a predesigned scheme or filing system. As an intellectual activity, classification includes document contents analysis and matching of its subject with the subject of some existing file. Besides the necessity of having good insight into the wide body of literature, an automatic system for classification of office documents requires considerable work related experience and practical knowledge. This type of experience and knowledge is usually case dependent and specific to a particular office, therefore creating a spectrum of different approaches. The challenge is to identify a standardized approach with a more generally applicable methodology which could be applied to various office classification systems.

There are a number of techniques developed by other fields that can be successfully utilized for automatic classification. Most of them are well elaborated, tested and reported in relevant literature. Systematic analysis and some generalization of available methodology suggests that there are at least three distinct methods and groups of available techniques applicable to the classification of office documents:

* statistical method

(techniques: word frequency, word weighting)

* linguistic method

(techniques: semantic, syntactic)

* artificial intelligence method

(techniques: expert systems, neural networks).

All of the techniques mentioned in this study can be used to some extent for the classification of office documents. However, the level of their overall suitability and applicability needs to be determined.

It is also important to keep in mind that the above mentioned approaches are influenced by philosophical, social, economic and other factors. All these factors and current trends play a significant role in the way researchers and scientists are tackling the world and their specific challenges. The global approaches in science have to be taken into consideration since they not only determine the methods and techniques, but also the current research trends. Various methods for classification of office documents are based on more general scientific methods and techniques. They form part of the generally available methodological apparatus with roots in different philosophical doctrines and beliefs. At the 5th International Study Conference on Classification Research, it was argued that the twentieth century could be characterized as the "age of analysis." There were three philosophical schools which ruled during this period: logical positivism, linguistic analysis, and systems analysis. Their paradigms directly influenced research in the area of classification where logical positivism could be traced in statistical approach techniques; linguistic analysis resulted in semantic and syntactic techniques; and systems analysis influenced the appearance of the artificial intelligence approach.

OBJECTIVES

Widespread use of word processors, electronic mail, facsimile messages, letters, memos, minutes and other documents created and kept in electronic form require a new approach to document handling and processing. Enthusiasm has been expressed with regard to document availability, ease of access, and their dissemination.(2) However, adequate methods and techniques for treating not only the form of these documents but also their contents are still fragmented and spread over various disciplines and different areas. It is believed that a comprehensive review of available methods and techniques would help researchers and potential users solve some of their methodological dilemmas.

With various objectives in mind different elements can be emphasized in a review of methods and techniques for automatic classification of office documents. During the literature review answers to the following questions were sought:

-- What is being classified in an office environment?

-- How is manual classification of office documents performed?

-- What is involved and what are the requirements for automatic classification?

-- What are the main methods available for automatic classification?

-- Which techniques could be successfully used for automatic classification?

-- Which would be the most appropriate techniques?

-- What are the main research trends?

-- Is the automatic classification of office documents feasible?

A wide range of topics and numerous literature on methods and techniques used for automatic classification, as well as literature on related subjects, was analyzed and organized into a framework aimed at fulfilling the following objectives which were established:

-- to identify main methods and underlying techniques that can be used for automatic classification

-- to examine basic assumptions and methodological tools used by main techniques

-- to determine the main direction of contemporary research in the area of automatic classification

-- to examine the role that artificial intelligence can play in automatic classification of office documents.

TERMINOLOGY

Common terminology is a necessary step for mutual understanding, since "it is only via proper definitions that one can recognize identical concepts.(3) In order to establish a coherent approach to the study of automatic classification of office documents the following terms had to be defined:

-- office documents

-- classification.

Office documents

Besides 'information,' 'document' is probably one of the most used terms in the area of records management, library and information science.

There are a number of constitutional elements in the definition of a document. Two are essential:

* document contents

(meaning of the document or the message conveyed)

* document form

(medium, carrier or message container).

It is not always easy in practice to make a clear distinction between these two elements. There is, more or less, general acceptance that documents represent recorded information. Interdependency of form and contents explains why a document is "frequently used interchangeably with record.(4)

Form, as one of the elements of the definition of a document, because of its tangible and changing nature, is more often studied than the document contents. Paper is the oldest and the most widespread document form. Any type of paper based record has been regarded as a document. Introduction of other media in records management has widened the spectrum of potential forms. The term document covers today "recorded information in any format."(5)

Harrod's Librarians' Glossary,(6) for example, adopts this approach of a document as a record which conveys information in any form. As some of the possible forms the Glossary lists images, audio recordings, maps, manuscripts, and tapes (video, audio, computer).

The same Glossary also offers the element of distinction. between office documents and other documents. Office documents are defined as the ones "drawn up or used in the course of an administrative or executive transaction."(7) Only documents created as part of an administrative or executive function qualify for the category of office documents.

Permanency is another traditional document characteristic. Whatever the form, information recorded has to remain usable and available for some period of time. This is the reason that some authors define documentary information as information that is recorded in some kind of permanent form, such as in written or printed materials. Regarding the medium itself, it is often elaborated that documents could be in paper, film, microfilm, magnetic media, or optical disk form.(8)

The Federal Records Act passed by the United States Congress gives very practical guidelines for assuring accurate and complete documentation of policies and transactions of the U.S. federal government, control of the quality and the quantity of records produced, and judicious preservation and disposal of records. According to the definition given in the Federal Records Act, the following elements qualify what is deemed to be a record:

-- any physical form (all books, papers, maps, photographs, machine-readable material)

-- made or received by an agency of the U.S. government

-- part of a public business transaction

-- evidence of organization, functions, policies, decisions, procedures, operations or other activities of the government, or the information value of the data in them.

The Federal Records Act also determined what should not be regarded as a record: library and museum material made or acquired solely for reference or exhibition, extra copies of documents preserved only for convenience of reference, stocks of publications.

This understanding of an office document, as elaborated in the Federal Records Act, was the one adopted in this study. The definition mentioned here will be used as a main guideline.

The battle to include electronic mail into the category of official U.S. government records was not a smooth one.(9) It took a number of court cases, decisions and rulings to include the E-mail into the category of official records. A mere print of the computer screen, for example, does not constitute printing of a record. To qualify as an official record, a paper print produced from an electronic mail system should include other information content from the electronic version, such as date of transmission, date of receipt, and distribution or routing.

The use of technology does not always solve all problems. In fact, in many instances it also creates new ones. According to some authors the traditional term 'record' is becoming meaningless in many automated paperless systems.(10) An electronically generated record may be an 'information view at a particular time' that simply lies in the eye of the beholder--a virtual record in an information virtual reality. Several contemporary technologies that are intended to reduce paper generation actually create the concept of a 'virtual document' by the way they organize information.

In their book on records management Johnson & Kallaus(11) use the terms records and documents interchangeably. They divide documents into four categories:

-- transaction documents

This is usually the largest part of document holdings in any office. It includes the day-to-day transactions of an organization (business forms, invoices, requisitions, purchase and sale orders, financial documents, legal contracts, personal records, etc.).

-- reference documents

This category includes all the documents which contain information needed to carry on the business over the years. These are documents referred to for decisions previously made, opinions, directives, plans. Correspondence, as well as reports and studies, are all part of this category.

-- external communications

Communications between organizations or between an organization and its individual customers, users; public notices.

-- internal communications

Communications between an organization and its employees (payroll records, bulletins, regulations), or among departments (inventories, inter office memorandums).

Some other authors(12) offer a more complex, but similar categorization of office document forms, such as:

-- action and non-action information

-- recurring and non-recurring information

-- internal and external information

-- historical and future information

-- documentary and non-documentary information.

Classification

Harrod's Librarians' Glossary(13) gives a comprehensive coverage of various views of classification. The following aspects of classification are defined:

(1) The act or action of arranging of things in logical order according to their degrees of likeness

(2) A written document, an actual scheme used for the arrangement of books and other material in a logical sequence according to subject or form

(3) Logic used or method applied for recognizing relations, generic or other, between items of information regardless of degree of hierarchy used, or whether those methods are applied in connection with traditional or computerized information systems.

An interesting observation can be made based on this definition of classification. There appear to be three main elements:

(1) physical action

(2) scheme

(3) logic.

These elements, if observed from an artificial intelligence angle, could be regarded as:

(1) user interface, use and achieved results (physical action)

(2) knowledge base (scheme)

(3) inference engine (logic or methods).

With this perspective based on the previous definition of classification, it could be concluded that the classification process is implicitly an expert system in action. In fact, it seems that classification is a type of manual document processing expert system. This interesting observation could in turn have a special significance for future identification and validation of computer based expert system technique for automatic classification of office documents.

Many definitions regard the essence of classification to be the grouping of, or arrangement of documents according to some adopted criteria. One of the definitions regards classification as an activity where documents are grouped according to their content and use.(14) This grouping is done according to a predefined "logical, systematic arrangement of material usually involving a set of symbols, numbers, or letters.(15)

Some records management specialists believe that "mentally determining the name of subject or number by which a specific record is to be filed(16) is the main element of classification. Others(17) regard 'subject' to be generally too ambiguous when applied to contents of documents or books and they prefer the term 'topic.' A topic designates the range of phenomena treated. It may be a single concept, but more often it consists of two or more related concepts.

Closely related to document classification is coding. However, a difference between classification and coding is often emphasized. Coding of documents is regarded to be the physical marking of an item, used to indicate in what classification it is to be stored.(18)

A distinction is sometimes made in literature among subject indexing, subject cataloguing and classification. Subject indexing is a process by which just parts of a wider bibliographic unit are identified. Based on this approach, compiling a back of the book index would be regarded as subject indexing. Using the same methodology, subject catalog is regarded as an activity where a bibliographic item as a whole is assigned a subject heading. Therefore, it is not a part that receives attention and gets its subject representation, it is rather a whole bibliographic item, i.e., book, report, etc. The previously mentioned physical activity of assigning a class number from some classification scheme represents, in fact, classification.

This approach is contested as being an "artificial, misleading and inconsistent" division.(19) As supporting evidence, database indexes such as the one in Chemical Abstracts are cited. Chemical Abstracts refers to complete books and reports as well as to their parts--chapters, articles, etc. Even libraries, in particular specialized ones, sometimes prepare a subject index for a whole bibliographic item as well as for its constitutional parts. Libraries usually call this analytical cataloguing or in the case of subjects--analytical subject cataloguing. This distinction regarding the term classification is even more confusing. Especially knowing that subject indexing may involve the use of some specific classification scheme or that a printed subject index might also follow a classification scheme. It is assumed that the reason for this confusing division is, in fact, the consequence of a failure to properly distinguish conceptual analysis from the process of translation or representation. Conceptual analysis and document representation are only two different stages of the process of indexing or subject cataloguing. The overall activity represents subject classification, i.e., forming classes of objects on the basis of their subject matter.

One of the popular records management textbooks(20) defines some of the principles of subject classification of office documents in the following manner. Classification is logical and standardized activity (from major to minor elements); practical activity (not academic); simple, functional, retention conscious activity (grouping similar subjects together); mutually exclusive activity (not ambiguous); and finally, flexible activity (allowing additions).

A somewhat different understanding of the term classification is adopted in the field of pattern recognition. Researchers from that field regard classification as a process of recognizing a pattern--an organized system of information. A recognized pattern is then placed into one or more predefined categories. If the pattern is successfully associated with a class, the pattern is believed to be properly recognized and classified.(21) Some systems designed to recognize patterns provide also a confidence estimate for their classification decision. It is, in fact, a measure of the degree to which the pattern-recognition system believes that the pattern information belongs to a specific class.

Existence of various definitions of classification requires a decision on a particular one to be followed in the research in order to avoid otherwise imminent misconceptions and misunderstandings. This review follows the definition given in A Basic Glossary for Archivists, Manuscript Curators and Records Managers,(22) where classification of office documents is regarded as the act of identifying documents or records in accordance with a predesigned filing system.

Since this review deals with methods and techniques for automatic classification, there is a need to shed some light not only on classification in general, but also on the concept of automatic classification. The review takes an approach that depending on the degree of involvement of human classifiers there are three levels or three types of classification:

-- manual classification

-- automated classification

-- automatic classification.

Manual classification, or classification in its classic form, is performed completely by a human classifier. Thus, involvement of computers in the process of decision making with regard to determining the actual class number is absent. It is important to mention also that using computers, for example, for retrieval of a previously entered document, checking or consultation of some data base or other available tool does not change the nature of manual classification. The important element is that the classification decision is made solely by the human classifier.

Automated classification, also called semi-automatic classification, goes a step further in using computers and creating computer programs as a type of aid or guideline.

Usually in automated classification systems, a computer requires a series of inputs, such as answers to the questions asked. Based on the answers, while using its internally built logic, it comes to a "conclusion' on the corresponding classification number. What is offered by the computer system is usually regarded as a suggestion which a human classifier can accept or can reject. A Classification of Office Documents (CLOD-X) expert system(23) is an example of such a system. CLOD-X operates in such a way that a number of questions is put to the classifier in an interdependent sequence (the previously given answer determines the subsequent question), allowing the system to induce the required classification number. It is then left up to the classifier either to accept the given suggestion or to come up with some other solution.

Automatic classification eliminates the use of human classifiers in the process of actual classification. Human classifiers, knowledge engineers and system programmers design and maintain the system, but they are not involved with the actual assignment of the individual classification number. System independently of a human classifier collects, normalizes, formats, sorts, validates, evaluates and automatically decides on the appropriate classification number. There are some requirements to be met if a system is to be considered as a fully automatic classification system. These requirements are:

-- electronic environment

Documents have to be available in electronic form in order to be further analyzed, processed and finally classified. They can be either documents originally created in electronic form or simply scanned paper based documents.

-- minimum human input

The very beginning of the classification process could be initiated by a human classifier, but all the rest has to be done independently by the system. Document analysis, rule firing, checking and assignment of the classification number have to be performed completely by the computer system.

-- reliability

Designed classification system has to be tested in the real office environment sufficiently enough to prove that it is capable of classifying actual documents at the success rate of at least 90 to 95 percent.

Besides the previous three required characteristics, a fully automatic classification system would have to have the following desirable characteristics:

-- expandability

System has to be expandable. It has to be capable of incorporating new files without additional rule programming.

-- self-learning capability

Previous experience and already classified documents should be used as references for faster and proper classification of new documents.

RESEARCH METHODOLOGY

The basic method used in this study was a review of the corresponding literature. A reading list was compiled through online search of DIALOG information system and LISA data base. In addition to this, the reading list also included literature cited in the most relevant publications or journal articles. Besides mostly articles, the literature surveyed included published books, some theses, research reports, etc.

The literature review concentrated on the subject of automatic classification of office documents. An attempt was made to cover the most relevant aspects and elements of this topic. This included in particular, the various methods, techniques and tools for automatic classification which are presently available. The emphasis was placed, however, on the literature which examines the possibilities of artificial intelligence solving the problem of automatic classification.

The following general categories were covered:

-- literature on records and office document management

Review of this literature concentrated mainly on the topics of the nature of office documents, their management and classification.

-- literature on automatic aspects of classification and cataloguing

This category included literature on the philosophical base and theoretical framework of automatic classification. Main emphasis was placed on various methodologies presently available and the achieved results. Benefits and difficulties of the use of computer based information technology in this area were also examined.

-- literature from other related areas

Topic areas such as natural language processing, automatic indexing, automatic abstracting, decision making, pattern recognition, document retrieval, and machine learning were taken into consideration, especially if they had some relevance to office documents classification.

The literature reviewed covered mostly items published during the last 10 years. It is only during this period that the development and widespread use of computers, especially personal computers, encouraged studies and research in this area. A few classic works published before this period have been included as well.

AVAILABLE METHODS AND TECHNIQUES

The reviewed literature suggests that there are three main methods, a variety of corresponding individual techniques, and their combinations thereof. The basic methods used for the classification of office documents are statistical, linguistic and artificial intelligence methods.

Statistical method

The basic premise of any statistical method is the possibility to index documents. Throughout history, manual indexing, or assignment indexing, was the main approach used by human indexers. This means that the indexer made a representation of a document using some type of controlled vocabulary. General index terms were not always present in the actual text but they were in some way associated with its contents. This type of indexing is very suitable for human indexers.

Extraction indexing (also known as derived or word indexing), a technique where only words or phrases present in the text were selected, was another type of indexing which was possible but less frequently used by human indexers because it was very time consuming and labour intensive. Computers happened to offer just a right tool for extraction indexing. Many authors believe that counting words, defining their frequency and position was ideal work for a "thinking machine," i.e., computer.

Word frequency

The main technique used in the statistical approach is word frequency analysis. One of the first authors to publish a paper in this area was H.P. Luhn.(24) Although his paper was entitled The Automatic Creation of Literature Abstracts, the main idea that he was experimenting with was not to produce abstracts, but rather extracts (indications of the document's subject). The method used was frequency of specific words. A computer program removed certain words such as articles and prepositions from the text using the "stop-list," counted the occurrences of remaining words, and then ranked the words according to some frequency criteria. The criteria was usually the highest or the lowest frequency number, or a possible higher or lower frequency number compared to a set occupance (cutoff) number. Based on the frequency of keywords within the sentence, each sentence was evaluated and the scores were compared. The first few highest scored sentences (depending on the desired length of abstract) were listed without any changes, making an automatic abstract. For a number of years to follow, this was the main method used in attempts to build systems for automatic abstracting.

There are other authors(250 besides Luhn that used the same method. They looked for groups of keywords clusters and calculated the value of each sentence based on their presence. Followers of the word frequency method also introduced the concept of relative versus absolute frequency of words in a document. Absolute word frequency would be the number of its occurrences in a particular document, while the relative number would be its appearance in a data base holding relevant documents. The selected subject terms in this case would be the words appearing in the document more than their relative frequency number.

There were many followers who believed that the statistical approach could bring the desired results in indexing, cataloguing or abstracting. For example, some used statistical calculation of keyword frequency in the document,(26) while others(27,28) used just simple keyword counts as the basis for computing sentence scores.

In an article on automatic abstraction,(29) Chris D. Paice provides a comprehensive review of techniques used for automatic abstracting. She lists seven different ways for evaluating sentence significance:

-- the frequency-keyword approach

-- the title-keyword method

-- the location method

-- syntactic criteria

-- the cue method

-- the indicator-phrase method

-- relational criteria.

Paice defines an abstract as "a concise summary of the central subject matter of a document."(30) According to her, there are two types of abstracts:

-- indicative abstracts

-- critical abstracts.

The present status of developments in this area, Paice concluded, makes construction of indicative, also known as informative or substantive abstracts, feasible. This indicative type of abstracts helps the reader decide whether it will be worthwhile to look at the full document, whereas the second type--critical abstract--contains useful information which eliminates the need to refer to the whole document. Her conclusion regarding critical and comparative types of abstracts was that it did not appear feasible that those abstracts could be automatically constructed at that time.

Experiments were done with the use of the statistical method for defining document subjects. In one of the attempts to define the subject content through title,(31) a hypothesis was tested that the more words there are in the title, the greater will be the likelihood that a Library of Congress subject heading will be in the title. Statistical probability has suggested that more words would increase the likelihood of a subject/keyword match. However, findings did not prove the hypothesis. In fact, fewer words in the title had a greater percent of match than did titles with more words.

Word weighting

Word weighting, another statistical technique, is very closely related to the word frequency counting. In fact, P.B. Baxendale(32) supplemented the word frequency criteria with an idea to use only the first and the last sentence in each paragraph. At the time when computer memory and speed were limiting factors, giving higher importance (weight) to some words enabled Baxendale to develop appropriate indexes more quickly using less resources. The main assumption of this technique was based on her study that the first sentence brings an 85 percent chance of a correct subject index hit, while the last sentence adds to it another 7 percent.

John M. Carroll(33) adopted Baxendale's approach that words occurring early in a document had a greater content-significance, He combined that approach with word frequency count to develop his system which can indicate "the most content-significant sentence in a block of text and extract a short list of content-significant words."(34)

Carroll used BASIC computer programming language as a development tool for this algorithm. In order to come up with the most important sentence on the page and the most important words, he defined four vectors:

-- text vector

to hold words and sentence delimiters and to reconstruct, as well as to display the sentence at the later stage

-- trunc vector

to hold first five characters of each word used to count words

-- work vector

to identify the number of times the word occurs

-- value vector

to calculate the weight of individual words.

After testing his system, achieved results allowed Carroll to conclude that "a simple computer algorithm can extract the most content-significant sentence from a short document as well if not better than a trained indexer can".(35)

Linguistic method

As opposed to the statistical method, which is more or less based on the mechanical processing of a document, linguistic methodology attempts to "understand" the document. Therefore, this method comes closer to human type processing where the key goal is to comprehend the information or the message conveyed by the document. The ultimate goal is not the words, but the actual subject or the concept which has to be recognized and used for further processing. It must be mentioned that there are still some naive concepts of subject,"(36) where the subject is always obvious and its identification never represents a problem. Fortunately, this idealistic and naive view is not that common since most of the researchers share the realist/materialist subject view. According to the realistic and the materialistic view things exist objectively and encompass objective properties.

There are two main techniques within the linguistic method. They are semantic and syntactic techniques.

Semantic technique

Semantic text analysis technique is well studied by a number of authors. The goal of this technique is to pinpoint the word(s) that give the best identification of a particular text or document. The main tools applied by this technique are location of keywords, use of controlled vocabulary, and use of indicator phrases.

Location of keywords. The main idea behind the keyword location approach is that the author has to use some key terms in order to express the general subject or idea of his document. The challenge is to come up with a methodology for the best identification of keywords. Besides the already mentioned word frequency, keywords could be determined in advance by a field expert allowing a computer program to successfully use them. A good review of techniques for keyword identification is listed at the end of this paper.(37)

The research in the area of semantic based keyword location made contributions to the following topics in particular:

-- keyword normalization

(stripping of prefixes and suffixes)

-- anaphoric substitutes

(words used to express/replace the keywords).

Many authors stressed in their studies the importance of anaphora.(38,39) They came up with an interesting conclusion that anaphora are used more often with main keywords than with some peripheral ones.

Use of controlled vocabulary. In classical library terminology a controlled vocabulary represents an authority list. "A controlled vocabulary is a subset of natural language, a limited set of terms that must be used to represent the subject matter of documents. The control vocabulary keeps only a small proportion of words, a few forms, and little or no grammar.(40) It is used for assigning terms to the document which is being processed. Controlled vocabularies control the use of synonyms, bring together the words which are related, and point out the homographs (words spelt in the same way but having different meanings).

There are three major types of controlled vocabularies:(41)

-- bibliographic classification schemes

-- list of subject headings

-- thesauri.

The main difference among them is the level of importance given to their alphabetic or systematic presentation. A list of subject headings brings the terms in alphabetical order, while thesauri, even though alphabetically ordered, give higher priority to hierarchical and associative relationships. Designing any of these control vocabulary tools

is a very complex and complicated job. For example, 300 panelists were involved in making the Engineers Joint Council(EJC) Thesaurus of Engineering and Scientific Terms.(42)

An interesting experiment was done by the U.S. Defense Documentation Center at the beginning of the 1970s. The idea was to use a controlled dictionary for machine-aided indexing. The project leaders(43) attempted to extract the words from the titles and abstracts and to compare them to what was named retrieval vocabulary--list of acceptable subject terms. It was reported that the achieved results by machine-aided indexing were at least comparable, if not better than the ones assigned by the human indexer.

Use of indicator phrase. The indicator phrase approach is based on the assumption that in order to emphasize the document subject, its important part or message, authors often use some specific phrases, for example "this paper studies...", or "the main subject is...", etc. Studying this, it is concluded that the indicator phrase could be used for the automatic generation of document abstracts.(44)

An experimental system for classification of office documents based on the semantic retrieval was developed at Olivetti.(45) The system describes and classifies documents by their semantic structure which provides access to abstract concepts contained in the document. The document's content is automatically analyzed using extended pattern matching which includes layout, logic, and content elements. The classification unit allocates the document to an appropriate predefined class type. Depending on the structure level of the tested document, the classification system performed differently. The classification of over 200 well-structured documents had the success rate of 90%, while the classification of less structured documents was only 70% successful.

An interesting semantic model for office documents classification and retrieval was developed at the Politechnical Institute in Milano, Italy.(46) The system assumed that there are three types of knowledge:

-- static knowledge

It incorporates the semantic information regarding the document types, the document contents, and the hierarchical relationships among document types and subtypes.

-- procedural knowledge

Knowledge of office procedures, their executors and triggers, that start or stop the execution of the procedure.

-- knowledge of the application domain

Description of links between the documents and set of regulations and laws which govern creation, use, retrieval, and retention of documents.

Syntactic technique

Syntactic analysis of a sentence attempts to identify grammatical types of words used (e.g., nouns, verbs, etc.) and their relation. This type of approach to indexing requires very close cooperation between information and language specialists.(47) Such cooperation being not too common, the literature dealing with syntactic analysis is less numerous than the one based on some statistical approach, for example.

An interesting linguistic model was proposed by N. Chomsky(48,49) who developed many of the basic notions of formal language theory. According to his model, a distinction should be made between surface structure and deep structure of language, in particular within sentence. Two sentences can use different words (surface structure) to convey the same meaning (deep structure). Using complex sentence parsing, Chomsky believed that syntactic analysis and transformation can provide the actual meaning of that particular sentence.

A well designed syntactic-recognition system has three main parts(50) performing the following functions:

-- signal to symbol transformer

Conversion of 'raw' signal or image data to symbolic form.

-- grammatical inference engine

Learning to recognize patterns.

-- syntactic organizer

Uses learned grammatical structures to recognize specific input sequences.

While the actual pattern-recognition systems usually combine more than one of the available tools, there are three basic tools applied in syntactic pattern-recognition systems:

-- prototype matching

-- parsing

-- grammatical inference.

The use of syntactic techniques to define the sentence semantics remains only a theoretical possibility yet to be practically proved.

Artificial intelligence method

Determining the suitability of a particular method to a specific problem or domain, such as the classification of office documents, is a challenging task. Artificial intelligence (AI), as one of the methods, is a powerful tool that can be of great help in many areas, but it is not a panacea. It is not necessarily suitable for every problem. Research and practice show us that numerous 'horror stories'(51) abound of disasters resulting from attempts to address problems using inappropriate technology.

The artificial intelligence method, as well as the previous two methods, has at its disposal a number of various techniques. The AI method is a relative newcomer to available methodologies when compared to other classic approaches. Artificial intelligence, in general, is regarded as an interdisciplinary field closely related to a number of other fields. The more important ones are mathematics, logic, computer science, linguistics, statistics, philosophy, psychology, and cognitive science.(52) The border between AI and non-AI techniques is still fuzzy,(53) but the number of available techniques is constantly growing. Most authors agree that there are two main and fundamentally different techniques imminent for artificial intelligence methodology. They are:

-- expert systems

-- neural networks.

However, there are still other authors who group the AI techniques in a very different manner. One catalogue of AI techniques consists of 256 techniques grouped into 19 categories of related techniques. Techniques were defined there as "algorithms, data (knowledge) formalisms, architectures, and methodological techniques which can be described in a precise clean way."(54) Here is the list of 19 categories and groups:

Automatic programming

Computer architecture

Data models

Expert systems

Game theory

Inference and reasoning

Knowledge representation

Learning

Logic programming

Natural language

Pattern recognition and image processing

Planning

Problem solving

Programming languages

Robotics

Search

Speech

Theorem proving

Vision.

In the part of the expert system group tools, this classification includes expert system shell, protocol analysis, and certainty factors. These tools are the most numerous, the most widely used and commercially available. The author of this classification had to admit that while compiling his catalogue, expert system shells threatened to swamp the other entries.

Expert systems

The use of expert systems methodology is more common than any other technique based on artificial intelligence. There are already a number of programs developed and successfully used, and there is also a variety of expert system shells commercially available. The expert system shell is a software package which incorporates all required elements such as inference engine, user interface, and a protocol for building a knowledge base and its rules. Therefore it enables users to develop their own expert applications without going too deeply into programming. In other words, expert system shell is a handy tool used to facilitate the development of an expert system.(55) It could be used for exploration, but also for implementation of the expert system technique. Easy access and good prospects for successful application are the main reasons for the expert system shell's popularity.

Expert systems usually consist of a knowledge base, inference engine and user interface. Many authors regard knowledge base to be the most important element.(56) The knowledge base holds the knowledge which is acquired either from available written sources or more often from an expert, through a process of knowledge elicitation. An

important way for building a knowledge base is observation. Acquisition of descriptive knowledge is regarded by many authors as the first step.(57) There are four types of functions that the knowledge engineer performs:(58)

-- knowledge acquisition

-- knowledge system design

-- knowledge programming

-- knowledge refinement.

The inference engine "decides" which heuristic search techniques are to be used to determine how the rules in the knowledge base are to be applied to the problem.(59) There are three main inference techniques used by expert systems. They are:

-- backward chaining

A process of working backward from observable manifestations to their probable explanatory cause.(60)

-- forward chaining

A new fact asserted to the system causes any rule whose premises match that fact to fire.

-- knowledge frames

A knowledge representation scheme that associates an object with a collection of features (e.g., facts, rules, defaults, and active values). Each feature is stored in a slot. A frame is the set of slots related to a specific object.(61) Closely related to them is also J. Ross Quinlan's Interactive Dichotomizer 3 (ID3) method which uses a process of induction from a set of examples to build a rule-based representation of decisionmaking in some domain.(62,63)

It is important to keep in mind that the critical moment before the actual use of any expert system is the decision whether the artificial intelligence approach is applicable. This decision should be based on a thorough analysis of the particular task and its specific subject area or domain. As already mentioned in the discussion on the terminology adapted for this study, classification in general seems to be a good candidate for the use of an expert system technique. In order to qualify as a suitable expert system candidate, classification has to fulfil some requirements. In particular, classification has to be an action that is done regularly and satisfactorily in everyday 'non-expert-system environment' by human experts. In other words, it should be possible with adequate knowledge and expertise, and after following given rules, to solve the classification problem. For example, diagnosing for engine malfunction or creating an optimal computer configuration are tasks which are regularly and satisfactorily performed by many experts in their respective areas. Thus, these two areas can be regarded as suitable candidates for application of expert systems technology. In fact, there are expert systems designed to solve problems in these domains presently available.

Contrary to this, there are problems where predicting an outcome is uncertain due to the nature of the problem. In these unsuitable areas, however proficient experts' background knowledge and experience might be, an expert system can bring nothing more than an educated guess. For instance, a knowledge-based application that could accurately predict stock prices or winners in a horse race would be extremely valuable. However, since no collection of humans has that type of expertise, a knowledge-based application built to solve these problems is not feasible.

According to some estimates(64) there were 2,200 expert systems installed by year 1992. Most of the expert systems are in the area of diagnosis where they were inferring system malfunctions from observable data. The second most popular area is interpretation where the expert systems are infeing situation description from available data. The third area is prescription where recommendations are being given to solutions of various system malfunctions.

Classification of office documents is a job regularly and satisfactorily performed in almost all organizations and institutions, so it meets the necessary requirements. In carrying out this activity regularly and satisfactorily, classification requires the existence of rules which allow experts to develop heuristics. Heuristic rules are usually the outcome of long working experience and sound judgment and are used as intellectual short-cuts for making more efficient and effective decisions in the domain area. It is necessary to have sustainable rules and heuristics that will govern problem-solving activities and allow knowledge engineers to design and construct required expert systems. Common sense, so readily available and taken for granted, is almost impossible to translate into computer programs. Computers operate more efficiently with a previously developed set of rules, such as the ones established for classification, indexing, or circulation, just some of the rules in the area of records management. It was suggested(65) that development of expert systems is possible only if the tasks are cognitive, easily understood and do not require common sense.

The classification of office documents also seems to fit this requirement. It is not a surprise to find numerous pages written on the expert systems technology as a valuable tool for solving various classifications tasks. There are many classification expert systems in operation today, which is a significant indicator by itself. Some authors(66) argue that the classification done by expert systems is a relatively straight forward task. That is, it is a process of deciding on a single choice among a set of predetermined, specifiable, and enumerable solutions. In other words, there is a high chance that an expert system project will succeed in classifying office documents.

There are also some opposite views. One of the studies(67) on classification and prediction methods which is derived from statistics, neural networks and expert systems, discusses also a possibility of machine learning. The conclusion reached there was not too encouraging. Expert systems were regarded as a good way for codifying 'rules of thumb,' but it was pointed out that the limiting factor was the complexity of knowledge representation which makes their use difficult.

Although artificial intelligence is a method by itself, it also borrows ideas and combines techniques from the other two previously mentioned methodological approaches. That was the case, for example, with the approach demonstrated in the project carried out by a group of researchers in 1989.(68) They have developed a rule-based system for exploring impedimenta to automating descriptive cataloguing from title pages. Obtained test results suggested that with a small set of rules it was possible to identify over 80% of the bibliographic fields on a random sample of title pages. Having problems with optical character recognition (OCR) devices, authors created machine readable files using Tex typographic language. Files were then interpreted using a parser program developed to extract space-delimited character-strings (tokens) and their

attributes (page location, coordination, type style, and type size). The parser program wrote tokens and attributes to a token file that served as the input to the expert system written in Prolog language. The second phase of processing involved building compound tokens--functionally meaningful units with a relevance to bibliographic fields. Compound tokens were created by adding tokens together until font style or size was changed, or until a vertical white space separation reached a specific threshold. Compounds were then searched for the occurrence of proper names (cities and publishing houses) using the substring matching procedure. The rule interpretation was the last step in the process of title based automatic cataloguing.

Another author(69) developed some specific tools and techniques necessary to perform automated conversion of natural language text into a structured knowledge base. This approach was, to say the least, very original. A chapter of a textbook on cardiovascular pathophysiology was taken out and converted into a structured knowledge base to be used as the domain knowledge base of a tutoring system. The technique used involved four stages:

-- preparation of text in which raw text from the chapter was converted into a form suitable for the LISP parser by removing incompatible data, correcting errors, numbering sentences, and converting all characters to upper case

-- parsing of text was done using Linguistic String Parser (LSP) developed at New York University with some vocabulary and grammar enhancements

-- conversion of parse trees to information. formats

-- conversion of information formats to knowledge frames.

The tool used for the last two phases was CLIPS--an artificial intelligence language developed at NASA's Johnson Space Center.

Somewhat similar approach was used in the TemplateFiller system developed at EDS Research in Albuquerque, New Mexico. It is a system which according to its authors "goes beyond text retrieval and categorization and simple methods of summarization to the actual extraction of information from the text."(70) The triggering element for the development of this system was a need to keep abreast of new products in the fast growing and rapidly changing computer industry. TemplateFiller reads articles from computer industry journals and automatically generates bulletins on new computer hardware and software products. This is achieved through a series of various interdependent processing steps. After downloading PC articles from the DIALOCT service, a Bayesian categorization system(71,72) is used for automatic identification of relevant articles. The preprocessing steps involve format normalization, recognition and labeling of domain specific semantic objects--potential slot fillers. This is followed by keyword-matching which discards individual sentences of no importance to further processing. The central point of the system is a generic linguistic processing which transforms the text into a representation showing underlying semantic relations between objects, including the potential slot values. At the end of the processing chain the template builder uses that representation to select values for slots in the template.

Particularly interesting is the research done on expert systems in the area of automatic decision making. A number of studies were carried out with the goal of finding the best approach when a number of different solutions is possible. "Computer Systems That Learn(73) describes different ways that computers can be used to make decisions without human intervention. Authors concentrate on learning systems that extract decision criteria from samples of solved cases available in a machine readable form. Their book is based on classification and prediction methods ranging from statistics, neural networks, machine learning and expert systems.

There are at least three areas of artificial intelligence in which decision making has a special interest and where some significant contributions for artificial intelligence in general were made. These areas are assumptions, conflicting objectives and preferences, and uncertainty.

Assumptions

A decision maker often does not have all the elements necessary for reaching the decision, so some assumptions or hypotheses have to be made. Researchers have so far established at least two distinct techniques for making assumptions:

-- a default assumption technique

With this somewhat troublesome type of assumption making, the first step is to establish that there is no "knowledge to the contrary" and then to make the assumption. However, some "knowledge to the contrary" can be discovered later which would require backtracking and changing of previously made assumptions. This backtracking is also known as truth maintenance or reason maintenance.(74) Since the list of assertions in such systems does not grow monotonically, expert systems using this technique are called non-monotonic.

-- a closed-world assumption technique

This type of assumption making is different from the previous one. The technique by its virtue assumes that all that is not known is false and it continues to build the system on that. Contexts, viewpoints, and hypotheticals are the terminology and apparatus used in systems based on this approach.

Conflicting objectives and preferences

This is another area that has attracted special attention in decision making, as well as in other areas which are contemplating the use of artificial intelligence techniques. A decision based on conflicting objectives and preferences is the one where there are a number of variables which might have conflicting repercussions on the actual decision. It is not the absolute magnitude of some value that matters as much as relative rankings of alternative variables or alternative courses of action.(75) For example, a selection of a university type of decision making can have some conflicting variables such as the level of tuition fees, distance from the residence, availability of free places and appropriate courses. A broad spectrum of research techniques was developed ranging from approaches with formal foundations, such as conjoint methodology, to the purely behavioral ones using protocol analysis.

One of the articles found on genetic algorithms(76) examines, for example, two types of decision models:

-- a model with compensatory rubs

In this model good features of an alternative can compensate for some bad features.

-- a model with noncompensatory rules

Since multiple attributes create complexity, decision maker simplifies the decision with a help of heuristic techniques. For example: a) elimination by aspects (EBA) in which choices that do not meet a threshold value are eliminated; b) lexicographic in which a decision is made based on the most important alternative; c) conjunctive technique where alternatives are eliminated if they do not pass a combination of thresholds; and d) phased EBA or compensatory where after using EBA technique, the compensatory rule is used to remaining alternatives.

This genetic algorithm approach begins with a superset of elementary decision operators which can make some good choice rules. The genetic algorithm searches to combine the operators in a manner that best fits the choice data. The operators are defined from common decision heuristics. For instance, a decision maker can look across alternatives or look across attributes; values can be compared to each other, maximum and minimum values may be noted, alternatives may be eliminated, or any other combination can be established until a decision is made.

Uncertainty

Decision making under uncertainty is an important topic in artificial intelligence, and it was studied by a number of researchers.(77)

-- Numeric assessments of likelihood

Classical probability theory, Baye's rule, statistical and subjective probabilistic techniques use the numerical assessments of likelihood approach. This approach inspired some further research in developing new techniques such as fusion, propagation and structuring.(78)

-- Probability updates

Certainty factors, theory of confirmation, and other nonprobabilistic techniques are based on the probability concept, where a degree of change in a system's belief based on new evidence is taken into consideration. Some authors argue that this non-probabilistic approach is in fact only a variation of the better established probabilistic approach.

-- Least-commitment approach The Dempster-Shafer mathematical theory of evidence(79) uses this approach. This probabilistic approach assigns belief to an individual hypothesis based on evidence. At the same time, it also assigns belief to some sets of hypotheses which evidence does not fully support and distinguish.

-- Plausibility

The plausibility approach is directly related to fuzzy logic. Fuzzy logic and fuzzy set theory are based on the premise that a description of an object or a situation from the real world is never precise. As Zadeh as put it, it is only plausible to a certain degree.(80) One of the books devoted to fuzzy logic and its founder Lofti Asker Zadeh stated that the "complex systems defy human comprehension and evade even definition."(81) The main tool for dealing with complexity and information flood successfully applied by the human brain is summarization. Through the process of constant summarization reducing "massive detail to chunks of perception,"(82) we manage to control and comprehend the vast amount of information regularly received by our senses. Once information is summarized we are able to go a step further and classify what was perceived. This enables us to classify the elegantly shaped Jaguar and the 'boxy' Volvo under the same category of "car." Classes concentrate on and highlight the common, while they dim down the unique details. In other words, precise is perceived in a fuzzy way.

-- Taxonomy organization of descriptions

This is a standard artificial intelligence approach used, for example, by a structural encoding technique. It uses a taxonomic organization of description and represents uncertainty by picking that node in taxonomy that subsumes all possible descriptions.(83)

-- Dialectical argumentation

Dialectical argumentation was used as a base for a theory of endorsements.(84) According to this theory, each piece of evidence is examined and elaborated, creating one set of arguments for and another set of arguments against.

As a summary of the issues regarding expert systems approach to modeling human work and decision making, the case of the n-Cube expert system(85) should be mentioned. It was an expert system for item classification using Universal Decimal Classification. During its development almost all of the above mentioned challenges were encountered. Unfortunately they were encountered as limitations. The limitations experienced by the n-Cube expert system included redundancy, conflict, specialization, ambiguity, and similarity. This was an obvious proof that ready-made solutions to at least some problems and limitations are required if an expert system is to be developed and implemented successfully. Otherwise, the amount of required work and questions to be solved would be just overwhelming.

Neural networks

Throughout human history man has tried to mimic nature and his own body when trying to develop some useful machine which was to replace some of his functions. The first attempts of building a "flying machine' resembled mechanical birds. Making a "thinking machine" was always a goal for many adventurous scientific minds. Neural networks is one of the attempts to design a thinking machine by mimicking the human body, in this case the human brain. The main difference between this attempt and previous human-like models of various machines is that neural networks promise success.

In neural networks a set of processing elements (nodes) modeling the brain's neuron cells is interconnected on multiple levels. Each processing node has both input and output capabilities, enabling them to interact with other processing elements by exciting or inhibiting them. Because of its multilevel connections, the impulses spread in the form of "spreading activation."(86) The whole process depends on a fairly complex system of modulations by various mathematical functions and algorithms.

The learning process in neural networks is based on the relative strengths of the node input connections and their modifications. Just one pass of input data can fix the

connections between the nodes, indicating that the neural network has learned the particular task. In practice, it is usually a number of trials and passes that is required before a task is considered set and the procedure acquired or learned. This learning is also known as neural network training. The learning process is one of the distinct characteristics of the neural networks. This characteristic was used in defining neural networks as networks of adaptable nodes which, through a process of learning from task examples, store experiential knowledge and make it available for use.(87)

There are two important factors which have facilitated the spread of neural networks' popularity:

-- significant developments in research on network types and characteristics

-- availability of new powerful computers.

This was followed by numerous successes achieved in demonstrating the ability of neural networks to deliver solutions to problems, particularly in the area of learning and pattern recognition.(88)

There are a number of neural network types developed. The most popular ones are:

-- Hopfield networks

-- Kohonen networks

-- Back propagation networks

-- N-tuple networks.

The main characteristics that make neural networks more human-like and differentiated from other techniques are:(89)

-- neural networks can handle incomplete data, partially incorrect data, and conflicting data while performing at the satisfactory level

-- they display spontaneous generalization and analogical reasoning

-- competing hypotheses about solutions to problems could be simultaneously checked

-- neural networks use "best match" feature

-- they can apply "fuzzy logic" especially important in classification where no one member of a class displays all the features characterizing that class

-- they benefit from "self-learning" using complex knowledge structures which humans are not always able to grasp, or may not even be aware of.

It needs to be emphasized that neural networks represent a new generation of knowledge-based systems, more advanced than the previous ones. Still, from the point of view of our general classification of three main methods available, neural networks are heavily dependent on the statistical method. Statistical methods are used to calculate activation levels and the modification of numeric weights. One of the drawbacks of this approach is the fact that the knowledge used by neural networks is not represented in an explicit and directly accessible form. For example, there are no modifiable sets of rules and the results obtained are difficult to examine.

Records management, library and information work are very suitable fields for the application of neural networks. Besides decision making, it is document classification and information retrieval that represent the best candidates. There are already some small but very significant experiments done in that area. For example, a neural network information retrieval system was constructed(90) based on only 407 documents indexed by 133 descriptors. Using a simple architecture, excellent retrieval results were demonstrated. This neural network system was able to locate documents which were relevant to the query request even though they did not contain the query term. This was possible because of connections established between various documents and index terms used in them. One document would activate another and so on, strengthening the index based interconnections along the way.

Another neural network system, called AIR (Adaptive Information Retrieval),(91) was developed using a similar approach and with the same goal of information retrieval. The AIR system went a few steps further. This system had integrated the ability to learn by modifying the connection strengths of linked nodes. Besides the documents and keywords, it also included a layer of author nodes making queries more flexible and producing better retrieval results.

CONCLUSIONS

During this review of available methods and techniques for automatic classification of office documents, a substantial amount of literature was consulted. The information gathered offered a solid base for drawing some valid conclusions. Pursuing the four objectives set out for this study, the following four conclusions can be made:

Firstly, there are three main methods available for classification of office documents: statistical, linguistic, and artificial intelligence methods. The statistical method starts from an assumption that techniques such as word frequency and word weighting could locate the subject and determine its classification elements such as the classification number, document routing, retention and so on. However successful the results of using this method might seem to be, this is still a very mechanical approach. Statistical techniques can become very sophisticated but they will always lack the basic prerequisite necessary for successful classification--understanding of the concept or the message conveyed by the particular document. The great value of the statistical method is its potential use by other methods, where the result of the statistical processing of a document can be used for further reasoning, document understanding and actual satisfactory classification.

The linguistic method goes a step further in attempting to understand the documents through its syntactic or semantic analysis. Still, the challenge for the linguistic method is to overcome its preoccupation with the meaning of single words, their clusters or sentences, or their grammatical structure. What is needed is a more general approach and understanding of the document as a whole. A good classification system has to understand the concept which is being conveyed by a particular document, not by some of its parts, however important those individual parts might be.

Artificial intelligence in comparison to other methods strives to go the furthest. Both expert systems and neural networks aim at gaining deeper understanding of a particular document and base their reasoning on that knowledge. Artificial intelligence systems are the most complicated ones to design. They require thorough knowledge of very specific methodology if they are to become successful. A very close and coordinated work between knowledge engineers and information/documentation specialists (e.g., records managers) is required.

Secondly, the methodology for application of main techniques is well established. This is of great importance to the actual implementation of automatic classification of office documents. Many research hours could be saved using already developed and tested techniques applied in different but related areas. Of special interest are indexing, abstracting, cataloguing, retrieval, natural language processing, and decision making. Borrowing and implementing techniques from the mentioned areas, adapting them to be suitable for automatic classification do not decrease the amount of significant development work which has to be done, but can substantially facilitate it.

Thirdly, expert systems are widely used. There is a wide spectrum of various expert systems presently being developed or already in operation. Efforts should be made to study their methodological foundations and problem solving strategies which are of direct interest to automatic classification of office documents. As a possible direction for future research, compiling an inventory of these expert systems could enable their more systematic analysis and comparison.

Fourthly, there seems to be enough evidence to conclude that expert systems is a promising technique for automatic classification. of office documents. This review of various methods and techniques suggests that building such an expert system would be feasible. What is needed is an experimental design, which could be used for live testing of real documents from an actual office setting. Only through real life situations and a number of case studies can wider conclusions, applicable to automatic classification of office documents as a whole, be drawn.

It is a great challenge to design a working system which can replace document classification--an area of human activity always regarded as highly intellectual. Basic methods and techniques are available, what is needed is their ingenious adaptation, development and application.

ENDNOTES

1. SVENONIUS, Elaine (1992). Classification: Prospects, Problems and Possibilities. In Classification Research for Knowledge Representation and Organization. Proceedings of the 5th International Study Conference on Classification Research. Amsterdam: FID

2. HAHN, Udo & REIMER, Ulrich (1988). Automatic Generation of Hypertext Knowledge Bases. Conference on Office Information Systems, March 23-25, 1988, Palo Alto, California. New York: ACM, p. 182-188

3. DAHLBERG, Ingetraut (1992). Knowledge Organization and Terminology: Philosophical and Linguistic Bases. International Classifier, Vol. 19, No. 2, 1992, p. 65-71

4. EVANS, Frank B., Harrison, Donald F.; Thompson, Edwin (compilers) (1974). A Basic Glossary for Archivists, Manuscript Curators and Records Managers. American Archivist, Vol. 37, No. 3, July 1974

5. GILL, Suzanne (1988). File Management and Information Retrieval Systems. USA: Libraries Unlimited, p. 192

6. HARROD'S LIBRARIANS' GLOSSARY (1990). Compiled by Ray Prytherch. Seventh Edition. UK: Gower

7. Ibid., p.204

8. SMITH III, Milburn D. (1986). Information and Records Management: A DecisionMaker's Guide to Systems Planning and Implementation. New York: Quorum Books

9. SKUPSKY, Donald S. (1994). The Law of Electronic Mail--The Impact of the White House Case on You] Records Management Quarterly, January, 1994, p. 32-40

10. PHILLIPS, John T. Jr. (1994). Virtual Records and Virtual Archives. Records Management Quarterly, January 1994, p. 42-45, 60

11. JOHNSON, Mina M. & KALLAUS, Norman (1982). Records Management. Third Edition. USA: K68 South-Western Publishing

12. ROBEK, Mary F. & BROWN, Gerald F., MAEDKE, Wilmer O. (1987). Information and Records Management. Third Edition. USA: Glencoe Publishing

13. HARROD'S LIBRARIANS' GLOSSARY (1990). Compiled by Ray Prytherch. Seventh Edition. UK: Gower, p. 197

14. SMITH III, Milburn D. (1986). Information and Records Management: A Decision-Maker's Guide to Systems Planning and Implementation. New York: Quorum Books

15. GILL, Suzanne (1988). File Management and Information Retrieval Systems. USA: Libraries Unlimited, p. 192

16. JOHNSON. Mina M. & KALLAUS, Nonnan (1982). Records Management. Third Edition. USA: K68 South-Western Publishing, p. 14

17. LANGRIDGE, Derek (1991). Classifying Knowledge Knowledge and Communication Essays on. the Information Chain. Edited by A.J. Meadows. London: Library Association Publishing, p. 1-18

18. JOHNSON, Mina M. & KALLAUS, Norman (1982). Records Management. Third Edition. USA: K68 South-Western Publishing

19. LANCASTER, F.W. (1990). Indexing and Abstracting in Theory and Practice. London: The Library Association, p. 15

20. ROBEK, Mary F. & BROWN, Gerald F.; MAEDKE, Wilmer O. (1987). Information and Records Management. Third Edition. USA: Glencoe Publishing

21. ROTHMAN, Peter (1992). Syntactic Pattern Recognition. AI Expert, October 1992, p. 41-51

22. EVANS, Frank B., Harrison, Donald F.; Thompson, Edwin (compilers) (1974). A Basic Glossary for Archivists, Manuscript Curators and Records Managers. American

Archivist, Vol. 37, No. 9, July 1974, p. 5

23. SAVIC, D. (1994). Designing an Expert System for Classification of Office Documents--A case study of CLOD-X. Records Management Quarterly, Vol. 28, No. 3, July 1994,p.20-29

24. LUHN, H.P. (195B). The Automatic Creation of Literature Abstracts. IBM Journal of Resource Development. Vol. 2, No. 2, p. 159-165

25. OSWALD, V.A. Jr. et al. (1959). Automatic Indexing and Abstracting of the Contents of Documents. Los Angeles, CA: Planning Research Co.

26. EDMUNDSON, H.P. (1969). New Methods in Automatic Abstracting. Journal of ACM, Vol. 16, No. 2, p. 264-285

27. Rath, G.J., RESNICK, A., SAVAGE R. (1961). The Formation of Abstract by the Selection of Sentences: Part I: Sentence Selection by Men and Machines. American Documentation, Vol. 12, No. 2, p. 139-141

28. EARL, L.L. (1970). Experiments in Automatic Extracting and Indexing. Information Storage and Retrieval, Vol. 6, No. 6, p. 313-334

29. PAICE, Chris D. (1990). Constructing Literature Abstracts by Computer: Techniques and Prospects. Information Processing & Management. Vol. 26, No. 1, p. 171-186
30. Ibid., p. 171

31. KELLER, Barbara (1992). Subject Content Through Title: A Masters Thesis Matching Study at Indiana State University. Cataloging & Classification Quarterly. Vol. 15 No. 3, p. 69-80

32. BAXENDALE, P.B. (1958). Machine-made Index for Technical Literature--An Experiment. IBM Journal of Research and Development, No. 2, 1958, p. 354-361

33. CARROLL, John M. (1981). Content Analysis as a Word-Processing Option. ACM SIGIR Forum 16, 1 (Summer, 1981), p. 126-129.

34. Ibid., p. 126

35. Ibid., p. 129

36. HJORLAND, Birger (1992). The Concept of 'Subject' in Information Science. Journal of Documentation. Vol. 48, No. 2 June, p. 172-200

37. EARL, L.L. (1970). Experiments in Automatic Extracting and Indexing. Information Storage and Retrieval, Vol. 6, No. 6, p. 313-334

38. BONZI, S. & E. Liddy (1988). The Use of Anaphoric Resolution for Document Description in Information Retrieval. Proceedings of SIGIR 88, Grenoble, June 1988

39. LIDDY E. et al (1987). A Study of Discourse Anaphora in Scientific Abstracts. Journal of the American Society for Information Science. Vol. 38, No. 4, p. 225-261

40. BELLARDO, Trudy (1991). Subject Indexing: An Introductory Guide. Special Library Association: Washington, D.C., USA, p. 20

41. LANCASTER, F.W. (1990). Indexing and Abstracting in Theory and Practice. London: The Library Association

42. BORKO, H. & BERNIER, C.L. (1978). Indexing Concepts and Methods, New York: Academic Press

43. KLINGBEL, P.H. (1970). The Future of Indexing and Retrieval Vocabularies. Alexandria, VA, Defense Documentation Center

44. PAICE, Chris D. (1981). The Automatic Generation of Literature Abstracts: an Approach Based on the Identification of Self-indicating Phrases. In R.N. Oddy et al, Information Retrieval Research. Butterworths, p. 172-191

45. EIRUND, Helmut & KREPLIN , Klaus (1988). Knowledge Baged Document Classification Supporting Integrated Document Handling. Conference on Office Information Systems, March 23-25, 1988, Palo Alto, California. New York: ACM, p 189-196

46. CELENTANO, Augusto, FUGINI, Maria Grazia & POZZI, Silvano (1990). Knowledge Based Retrieval of Office Documents. 13th International Conference on Research and Development in Information Retrieval. Organized by ULB Brussels, Belgium 5-7 September 1990. ACM SIGIR '90. Edited by Jean-Luc Vidick, p. 241-253

47. SPAREK-JONES, K. & KAY, M. (1973). Linguistics and Information Science. Academic Press, New York

48. CHOMSKY, N. (1956). Three Models for the Description of Language. IRE (Institute of Radio Engineers.) Transactions on Information Theory, IT-2, Vol. 2, No. 1, p 111124

49. CHOMSKY, N. (1957). Syntactic Structures, Mouton and Co. The Hague

50. ROTHMAN, Peter (1992). Syntactic Pattern Recognition. AI Expert, October 1992, p.41-51

51. WALTERS, J. & NIELSEN N.R. (1988). Crafting Knowledge-Based Systems: Expert Systems Made Easy Realistic. USA: John Wiley & Sons, Inc.

52. WECKERT, J. & MCDONALD, C. (1992). Artificial Intelligence, Knowledge Systems, and the Future Library. Library HI-TECH. Vol. 10, No. 1-2 (Consecutive issues 37-38), p. 7-13

53. BUNDY, Alan (1990) Ed. Catalogue of Artificial Intelligence Techniques. Third, revised edition. Springer-Verlag, Berlin

54. Ibid., p. V

55. IGNIZIO, James P. (1991). Introduction to Expert Systems: the Development and Implementation of Rule-Based Expert Systems. New York: McGraw-Hill, Inc.

56. Larichev, Oleg I. (1992). A New Approach to the Solution of Expert Classification Problems. Current Developments in Knowledge Acquisition--EKAW '92. 6th European Knowledge Acquisition Workshop, Heidelberg and Kaiserslautern, Germany, May 18-22, 1992. Berlin: Springer-Verlag, p. 283-297

57. MANAGO, Michael & CONRUYT, Noel (1992). Acquiring Descriptive Knowledge for Classification and Identification. Current Developments in Knowledge Acquisition--EKAW '92. 6th European Knowledge Acquisition Workshop, Heidelberg and Kaiserslautern, Germany, May 18-22, 1992. Berlin: Springer-Verlag, p. 392-405

58. HAYES-ROTH, Frederick (1992). Knowledge Systems: An Introduction. Library HI-TECH. Vol. 10, No. 1-2 (Consecutive issues 37-38), p. 15-29

59. MISHKOFF, Henry C. (1986). Understanding Artificial Intelligence. Texas: Radio Shack

60. SZOLOVITS, Peter (1987). Expert Systems Tools and Techniques: Past, Present and Future. In AI in the 1980s and Beyond: An MIT Survey, edited by W. Eric L. Grimson and Ramesh S. Patil. The MIT Press, Cambridge, Massachusetts

61. HARMON, Paul & King, David (1985). Expert Systems: Artificial Intelligence in Business. New York: A Wiley Press Book

62. QUINLAN, J. Ross (1979). Discovering rules by induction from large collections of examples. In Expert Systems in the MicroElectronic Age, Edited by D. Mitchie. Edinburgh: Edinburgh University Press

63. QUINLAN, J. Ross (1986). Induction of Decision Trees. Machine Learning, No. 1, p. 81-106

64. DURKIN, John (1994). Expert Systems: An Overview of the Field. PC-AI, January/February 1994, p. 37-39

65. WATERMAN, D.A. (1986). A Guide to Expert Systems: Reading. USA: AddisonWesley

66. CLANCEY, William J. (1984). Classification Problem Solving. Proceedings of the National Conference on Artificial Intelligence. August 6-10, 1984 at University of Texas at Austin, p. 49-55. Classification Theory in the Computer Age: Conversations Across the Disciplines. Proceedings from the Conference, November 18-19, 1988, Albany, New York. University of Albany

67. WEISS, Sholom M. & KULIKOWSKI, Casimir A. (1991) Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems. USA: Morgan Kaufmann Publishers

68. WEIBEL, Stuart, OSKINS, Michael & VIZINE-GOETZ, Diane (1989). Automated Title Page Cataloging: A Feasibility Study. Information Processing and Management, Vol. 25 No. 2, p. 187-203

69. MAYER, Glenn Norman (1992). Creating a Structured Knowledge Base by Parsing Natural Language Text (PhD thesis). Illinois Institute of Technology, Chicago

70. SHULDBERG, Kelly H. (1993). Distilling Information from Text: The EDS TemplateFiller System. Journal of the American Society for Information Science, 44(9), p. 493507

71. HILL, J. & SCHNEDAR, M. (1992). Bayesian Procedures for Automatically Categorizing Text Documents (Technical Paper). Albuquerque, New Mexico

72. MOSTELLER, F., & Wallace, D. (1964). Applied Bayesian and Classical Inference: The Case of the Federalist Papers. New York: Springer-Verlag

73. WEISS, Sholom M. B KULIKOWSKI, Casimir A. (1991). Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems. USA: Morgan Kaufmann Publishers

74. DOYLE, J. (1980). A Model for Deliberation, Action, and Introspection. Artificial Intelligence, No. 12, p. 231-272

75. WELLMAN, M.P. (1985). Reasoning about Preference Models. MIT Laboratory for Computer Science TR 340

76. OLIVER, Jim (1994). Finding Decision Rules with Genetic Algorithms. AI Expert, March, 1994, p. 33-39

77. KANAL, L.N. & LEMMER, J.F. (1986). Uncertain in Artificial Intelligence. North Holland

78. PEARL, J. (1986). Fusion, Propagation and Structuring in Belief Networks. Artificial Intelligence, No. 29, p. 893-899

79. SHAFER, G. (1976). A Mathematical Theory of Evidence. Princeton University Press

80. ZADEH, L.A. (1978). Fuzzy Sets as a Basis for a Theory of Possibility. Fuzzy Sets and Systems, Vol. 1, p. 3-28

81. MCNIEL, Daniel & FREIBERGER, Paul (1993). Fuzzy Logic. New York: Touchstone, p. 16

82. Ibid., p. 43

83. SZOLOVITS, Peter (1987). Expert Systems Tools and Techniques: Past, Present and Future. In AI in. the 1980s and Beyond: An MIT Survey, edited by W. Eric L. Grimson and Ramesh S. Patil. The MIT Press, Cambridge, Massachusetts

84. DOYLE, J. (1980). A Model for Deliberation, Action, and Introspection. Artificial Intelligence, No. 12, p. 231-272

85. COSGROVE, S.J. & WEIMANN, J.M. (1992). Expert System Technology applied to Item Classification. Library HI-TECH. Vol. 10, No. 1-2 (Consecutive issues 37-38),

86. FORD, Nigel (1991). Expert Systems and Artificial Intelligence: An Information Manager's Guide. London: Library Association Publishing

87. ALEKSANDER, I. & MORTON, H., (1990). An Introduction to Neural Computing. Chapman and Hall

88. WARWICK, K. (1992). Neural Networks: An Introduction. In Neural Networks for Control and Systems. Edited by K. Warwick; G.W. Irwin and K.J. Hunt. London: Peter Peregrinus

89. FORD, Nigel(1991). Expert Systems and Artificial Intelligence: An Information Manager's Guide. London: Library Association Publishing

90. MOZER, M.C. (1984). Inductive Information Retrieval Using Parallel Distributed Computation. Institute for Cognitive Science, University of California, San Diego. ICS Report No. 8406

91. BELEW, R.K. (1989). Adaptive Information Retrieval: Using a Connectionist Representation to Retrieve and Learn about Documents. In Proceedings of the Twelfth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval. Cambridge, Massachusetts, USA, June 25-28 1989. Edited by Belkin, N.J. and Van Rijsbergen, C.J. New York: Association for Computing Machinery

Author: Dobrica Savic holds an M.Phil. degree in library and information science from Loughborough University of Technology, as well as a B.A. and M.A. in International Relations from Belgrade University. He has over 16 years of working and consulting experience in the area of information and documentation. He spent the last ten years with various United Nations agencies working in archives, registries, libraries and documentation centres.