Engineering the production of meta-information: the abstracting concern

Maria Pinto

Department of Information Science, University of Granada, Granada, Spain

Received 28 December 2002
Revised 30 March 2003

Abstract.

In order to improve the automatic production of meta-information in the abstracting field, an essential starting point is the exposition of the current state of the art. At the level of content, three significantly different types of procedure stand out, depending on the document structure in question: extracting, rhetorical summarizing and cognitive summarizing. In addition, reticular and graphic models of information representation, much more appropriate to digital environments, offer a complementary method. In all cases, prior definition of the domain, with its specific documents and actors, is needed. However, the low quality of the product derived from full automation (extract and summaries), above all lacking in coherence, led us to the concept of partial automation, a hybrid man–machine methodology that, at least for the time being, seems to be the best solution for the abstract and abstracting problem.

Keywords: automatic abstracting; natural language processing; semantic analysis; metadata; user needs

Correspondence to: M. Pinto, Department of Information Science, 18071 University of Granada, Granada, Spain. E-mail: mpinto@ugr.es

1. Introduction

Considering the increasing technological dependency on the processes of digital data storage and exchange and the need for a universal data management system, the concept of metadata appears to be both widely introduced yet vaguely defined. From a global standpoint, metadata (data-about-data) is primarily a basic resource descriptor that facilitates identification and retrieval [1] when an increasing number of documents are produced solely in digital format [2]. In relation to resources it has been suggested that simple metadata could be useful for location, richer metadata for the selection, and the richest metadata for evaluation and analysis [3]. A basic component of the so-called metadata will be the abstract, in view of its capacity to represent content and structure of all kinds of resources.

However, metadata increasingly, but not exclusively, refers to machine-understandable information, and consequently often to machine-produced information. The problem is that, despite the amazing growth of documents in electronic format and the diffusion of sophisticated self-editing systems, most abstracts are still composed manually today and cannot easily form part of the metadata systems. Therefore, automatic abstracting has become a primary need. Aside from all the human energy saved, there would be huge savings in time, with automated processes making the source document and the metadata available simultaneously. Assuming the computer to be capable of executing non-ambiguous tasks, as should be the case in preparing or accessing abstracts, the big problem of automatic abstracting is the understanding of information to the

extent that such understanding and further processing can add value.

If the computer can efficiently handle lexical-syntactic structures, and even – albeit with more difficulty – logical-syntactic ones, the greatest challenge remaining is the structuring and understanding of those psycho-sociological and pragmatic elements that occur in texts and can be understood only by people. At the present time, computers can manage information but not knowledge, a higher activity that belong exclusively to humans. Restricted to the layer of the information, each automated system relies on particular algorithms with more or less sophisticated combinations of statistical and linguistic components. While the computer can easily extract significant phrases from the document, the problems begin at the stage of producing authentic abstracts. There are continuing complaints about the low quality of the derived automatic product, especially in view of the high cost of that automation.

For each stage of abstracting – selection, interpretation and production – the difficulty in computer programs is different. Selective operations are fairly easily taken over by a machine and, since the syntax of any language is an extremely precise system, the computer can operate with lexical units and grammatical rules once the algorithms and the corresponding programs have been established. Difficulties arise when we move on to the interpretation stage, because of the semantic problems of ambiguity and imprecision. We might well imagine that, by throwing semantic ballast overboard, transforming the logical-semantic structures of the text into simply logical ones, the problem would disappear. An interpretative force is needed to transform contextual information into *personal* knowledge and that capacity, at least for the moment, belongs exclusively to the human condition. Nevertheless, information science research, carried out in conjunction with advances in cognitive psychology, linguistics and logic, could bring significant progress in the near future.

A pioneer in automating the production of abstracts was Luhn [4], who designed, in the late 1950s a procedure based on the distinction of words with a high conceptual content. Once the sentences with the highest number of significant words are identified, the program calculates the significance factor of each sentence, determined by the number of semantic clusters and their value in terms of meaningful words. The resulting abstract is based on the sentences with the highest significance factor. Along similar lines, Mathis-Rush [5] outlined the following sequence for an automated abstracting system: the reading of the source document; the application of a series of rules of selection and transformation; the construction of the abstract; editing and printing.

Since the 1960s, considerable effort has been made to design automated abstracting methods by means of natural language processing (NLP), a powerful technology combining an array of computer techniques that imitate human language procedures. At present, NLP is an emerging sector of computational linguistics where linguistic theories co-exist with automatic text analysers, parsers, and socio-pragmatic models of discourse. This linguistic and computational integration has led to a large number of automation projects, such as translation, information analysis-retrieval and speech recognition.

As humans, we extract meaning from all levels (morphological, syntactic, semantic and pragmatic) of the synchronic language model [6]: morphological analysis was used in primitive search techniques for segmenting words. Morphological analysers can gather an inventory of the language units, as well as complementary grammatical information (prefixes, roots, suffixes, etc.) and a series of rules for production. Despite paying attention to the most important aspects of the words (analysis of the flexive, derivative and compound forms), it does not eliminate the ambiguity of natural language sentences. On the other hand, syntax studies the structural regularities of the sentences, provides means of reducing ambiguity, and is the basis of a compositional focus for interpretation. The automatic syntactic analysers (parsers) identify the elements of the sentence as well as their relationships, especially anaphora (using a pronoun instead of repeating a word), cataphora (the use of a grammatical substitute that has the same reference as a following word or phrase) and deixis (aspects of a communication whose interpretation depends on knowledge of the context in which the communication exists) [7]. At this syntactic level, automatic 'stoplists' have been developed to eliminate non-substantive words such as adverbs, articles and pronouns. 'Stemming' processes allow the detection of words appearing in different forms (plural, verb tenses) and reduce all their variants to the root or standard lexical form.

Through semantic analysers the computer can assign some meaning to the structures previously identified by the syntactic analyser. The knowledge in the NLP semantic network represents a system of complex conceptual structures made up of combinations and interrelations of simple concepts. Two kinds of semantic parser can be distinguished:

bottom-up, or full parsing, processes every word and consequently a complete lexical, syntactic, semantic and discourse knowledge is needed; t*op-down*, or partial parsing expectation-driven, based on knowledge of the domain in question, is more tolerant of unknown words or grammatical lapses and ignores many of the complexities of the language [8]. The applications show a definite trend away from systems that rely heavily upon knowledge of the subject domain towards systems that, alongside domain knowledge, incorporate knowledge about discourse structures. In any case, one of the major problems with semantic analysers is that strong bases of knowledge are required before moving on to a possible interpretation. A significant example is the powerful UMLS ontology, created to represent and index biomedical documents.

However, the comprehension of discourse cannot be reduced to a set of operations for codifying and decodifying symbolic messages. Instead, it must be integrated into communicative processes, accounting not only for the different knowledge bases of receivers and senders, but also for the documentary objectives and the subtle role of intentions and expectations. Here, at the pragmatic level, the concept of context plays a special role, since we will not be able to develop effective summarizing systems unless proper attention is paid to the three distinguishable context factors: input, purpose and output [9]. Context analysis plays a critical role in helping us to select useful and feasible directions for research into summarizing. This pragmatic line was taken in the work of Salton [10], concerned with the automatic transformation of texts. The reasoning behind this approach is that the output of text processing should be delivered in the form of passages of natural language text, because users are less likely to accept obscurely formal responses. Systems for automatically generating texts should begin by extracting a number of relevant parts from a knowledge base. Although this can be performed easily in restricted systems, the selection of relevant data is a much more intricate process in most other situations. Once the fragments of the knowledge base are identified, a problem-solving module selects a particular profile and style for the meta text, keeping in mind the objectives that must be fulfilled, and textual prescriptions are given accordingly. A discourse-organizing module relates the fragments of available information into well-formed natural language sentences on the basis of the importance of their terms. Finally, a sentence-generating module puts the text into paragraph form.

2. New roles of abstracts and other forms of representation in digital environments

We have to recognize that deep differences among structures, forms and contents of e-documents, if compared with their corresponding traditional ones, have led to profound changes in the information representation systems and consequently, in abstracting and abstracts.

Armstrong and Wheatly [11] conducted an experimental study based on the selection of two representative samples of abstracts from generic search engines and from online databases, comparing and analysing both the characteristics of contents (representativity, authority, depth, source) and their physical aspects (legibility, presentation, extension). This study revealed substantial differences between both kinds of abstracts as far as size, contents and format were concerned. For instance, the conventional expository form consisting of one or two paragraphs with natural language links between sentences is being replaced by the structured abstract, based on the rhetoric of scientific discourse. The abstracts of the search engines are briefer than those of databases and, furthermore, their short sentences and paragraphs make them very readable, despite their evident shortcomings, such as limited representativity. Contrary to the perception that, with the expansion of mechanized information and the evolution of the Internet, the new technologies would provide a panacea for recording all the new modes of digital information, the effectiveness of the systems is still less than satisfactory. This is because the end-user cannot know how the retrieval system is performing, and he or she must also grapple with increasingly complex search routines. Consequently, a sound knowledge of sources and user needs is necessary [12].

The exponential growth of web search engines, with their primitive algorithms, has failed to solve the problems of documentary search and retrieval. Some of these engines seem to have been developed in response to a real need but with no recognition of the history of the abstract as a tool developed over the years to satisfy just such needs [13]. As a tool for searching and retrieval, the abstract continues to be a practical need in the digital scene, although some substantial changes would enhance its usefulness in the new environment. Originally conceived to provide a solid analysis of documents in the early days of information systems, they are also vital tools of retrieval. An abstract is a rich semantic tool, capable of being well structured and

adapted to current search settings. Although designed to fit the need of information systems that had little in common with today's full-text systems, their basic concepts still prevail, and the problems surrounding their use have more to do with performance than with the appropriateness of their conceptual roots. Thus, abstracts could be the basis for an extensive semantic network that refines current intelligent retrieval systems. Ideally, a *hyperabstract* must incorporate, besides the basic semantic information, all kinds of links to other related abstracts and/or documents as well as facilities for browsing.

As a means for exploring resources, the utility of Internet abstracts has increased. Taking into account that most web users are not information handling experts but simply curious people with limited skills, they may increasingly use the abstract as a guide for accessing information. However, the lack of formally structured information continues to pose a problem. Because search engines automatically analyse web pages without any distinction of the location of words, sentences and paragraphs, an abstract tag or metadata would be very helpful. So the concept of *metadata* is the potential key for universal information management, considering it has the capacity to carry information that includes not only the content, but also the context and the links of digital documents [14]. Assuming a flat space of unrelated resources [15], one of the main challenges of the digital environment is the structuring of such resources, and consequently of the corresponding metadata, among which the abstract is an outstanding representative.

Thinking of the new objectives and dimensions of the electronic abstract in the context of the Internet, the proposed model of an abstract in ISO Standard 214–1976 is clearly outdated, and a new and richer taxonomy should be established [16]. The fact is that the growth of information has not always been accompanied by an increasing capacity to secure access to it. The thematic variety and the structural heterogeneity of electronic texts call for automatic processing that is still in need of research. The scant or random structuring of texts and the deficient methods for automatically building great hypertextual structures complicate the panorama even more. *Intratextual* links among mutually related parts of a text have been shown to be feasible – connecting specific paragraphs with others that have a similar content. Similarly, *intertextual* relationships could be discerned with other texts of the collection that share information. Additionally, in view of this richness of structures, the abstracts allow a multi-layered representation of documents [17]. Thus, we arrive at the concept of *interactive microtext*, a form of open reference, with vaguely defined limits and whose most important feature is the capacity for hypertextual links within a collection. Extracts, summaries and abstracts would form part of this new documentary category.

Much has changed since Borko and Bernier's [18] summary of the main contributions to automatic abstracting. In view of the proliferation of models and methods over recent years, it is very hard to classify automatic abstracting systems. We shall nonetheless offer a state of the art, with two major methodological clusters: *extracting methods*, from the surface structure; and *summarizing methods*, from the rhetorical/deep structures. In addition, the graphic and relational methods are particularly promising.

3. Extracting methods

Some relatively useful procedures for documentary representation are based on the extraction of words, sentences or significant paragraphs from the source text. Paice [19] speaks of the following statistical methods for automatic extracting:

(1) *Frequency-keyword* – the keywords expressing the central topic are selected, taking into account their frequency of appearance in the sentence, and considering the terms with medium frequency as the most adequate.
(2) *Title-keyword* – similarly, key words from the title and heading of the document are selected.
(3) *Location* – the sentences are given a numerical value based on their situation in the text (beginning/end of paragraph, beginning/end of document, immediately after a heading) and the more valuable are selected.
(4) *Cue words* – substantive words that occur frequently in a given document, yet are rare in the collection as a whole, are identified [20]. The sentences that contain groups of these words should prove to be characteristic of the document. The words with the highest scoring are designated as theme words, and the sentences are given a score according to a weighted count of the theme words that they contain [21].
(5) *Indicator phrase* – some expressions such as 'the purpose of the study is', 'the final conclusions of the work are' or 'the findings of our research show' are detected and used as content indicators.

(6) *Relational criteria* – significant information is extracted with a view to a semantic representation of the document.

A study carried out by IBM for the US Army, known as ASCI-Matic[22], used similar procedures to those proposed by Luhn, although with a further elaboration of the sentences sampling techniques and the classification of the documents to be abstracted. Documents with many non-substantive words, or with extremely long sentences, were given special treatment. Although this method is an improvement on Luhn's, with respect to the density of representative words in a single sentence, the algorithms required are very complex.

Edmundson [23] puts forward a mathematical method that focuses on four modes of sentence selection: *key words*, *cue*, *title* and *location*. The Luhn criterion of frequency serves to identify the key words. The cue refers to lists of words classified according to their significance as 'bonus words', which are positive or meaningful, 'stigma words', which carry little weight, or 'null words' irrelevant for sentence selection. The title method gives special weight to sentences using words that appear in the title or subheadings. Finally, the location method assumes that certain headings will be followed by particular and meaningful sentences. Experiments eventually showed that the cue–title–emplacement procedures were more effective than those of the key words, and these were dropped from automation. It was also suggested that syntactic and semantic structures should be taken into account, as well as the statistical variables.

One important contribution to abstracting came from indexing criteria, which could be applied to the selection of sentences with the most representative words [24]. To determine the significant words in the context of the document, the most frequent are identified and the adjacent words noted and classified as a 'multiterm'. The next step is to identify, order and select the phrases that contain most multiterms – in view of the desired extension of the abstract.

Rush *et al.* [25] conceived an abstracting process with a set of rules for selecting meaningful phrases and a word control list (WCL). The WCL, like a small dictionary, orders words and phrases alphabetically, and assigns each entry a semantic weight and a syntactic value. This combination, evoking the linguistic dichotomy grammar/dictionary, can be used to create indicative abstracts. The existence of chains of significant words incorporated into the dictionary could serve to eliminate a high percentage of empty sentences. However, the abstracts thus obtained were of poor quality.

Another automatic application for extracting information is the ANES (automated news extracting system) [26], which works exclusively within the field of journalism. Its statistical/heuristic analyses allow determination of the document marker, calculating the term's frequency and corresponding weight if the relevance of a term is proportional to its frequency in a given document, and inversely proportional to the number of documents in which it appears. To generate the extract, ANES chooses sentences that contain the previously selected word markers/identifiers. The coherence of extracts generated by ANES – between 60 and 250 words – compared with that of the source texts, was judged as being of only medium quality. As in other cases, we can observe that the information found within the initial sections and sentences of texts tends to be the richest in significant information.

Barzilay [27] proposes a summary generation system according to an algorithm that identifies lexical chains within the text. The full sentences extracted from the source text are those having lexical chains with extra-strong, strong or medium-strong relations.

Nevertheless, these isolated investigations with weak and questionable scientific bases offer few perspectives for progress. A more rigorous line of research can be found in Knowledge Discovery, a methodology for extracting information from a certain type of regularity appearing in the data. Our opinion is that knowledge is too ambitious a word, and it would be much more appropriate to speak of 'information discovery'. Taking into account that a document is always within a collection, the processes involved in KD take place in several stages. First there is pre-processing, whereby the document is transformed into a sequence of words after a process of reduction. Subsequent steps are the discovery of the maximum sequences of frequent words and the classification and ordering of the sequences of words assigned to each document based on a set of indicators (e.g. longitude, frequency, stability). Finally, the system has to determine the usefulness of the information discovered. The automatic extraction of the maximum frequent word sequences [28] in a group of documents allows a sequence of words to be found that is frequent (above a predetermined threshold of frequency, n) in one collection. The representation of a document as a group of sequences of frequent words (or a set of descriptors) can be useful in enabling quick responses to user queries, easing associations of close words and hypertext links to describe related documents. We

must emphasize the key role of Knowledge Discovery in today's information systems for developing a new array of tools and techniques to extract information from databases.

As we await the development of more intelligent expert systems, sentence extraction is another effective method of generating extracts. In order to obtain coherent abstracts from linguistic and logical standpoints, Mathis [29] operates with the selected sentences. For instance, if a sentence considered adequate for the abstract requires an antecedent to make sense, the three preceding sentences are examined to determine whether they should also be included. If not, the selected sentence is re-written so that it makes sense on its own. The next step, following a more recent trend, is the extraction of paragraphs instead of sentences, but these products do not have the expected coherence or comprehensiveness, and they still depend largely on interpretative repertoires.

4. Summarizing methods

Summary is the halfway point between extract and abstract, adding some creativity to the single extractive methodology. However, it does not have the status of an authentic abstract, especially as far as it most prized quality – coherence – is concerned, an attribute that is difficult to achieve automatically. While the physical or surface structure of the source is used in constructing extracts, its rhetorical and cognitive structures aid the production of summaries. Therefore the implemented research can be grouped around two trends: rhetorical and cognitive.

4.1. Rhetorical summaries

The rhetorical structuring of documentary sources establishes a cognitive script that enhances their informative capacity. The rhetorical structure of the automatic summary can be derived either from its equivalent in the source document, or from frames previously set up.

From the source document. Ono [30] describes a system for automatically summarizing expository texts based on the extraction or pruning of the rhetorical structure. The natural order of the sentences' relevance can then be seen, without affecting the semantic structure. Once the discourse structures are detected through surface connectives, and the sentences of the future abstract are determined, the system alternately orders these phrases and the connectives from which the semantic relationships were extracted. The resulting semantic structure is pruned depending on the abstract length. Marcu [31] applies empirical methods to demonstrate that there is a strong correlation between the trunk of the rhetorical structure trees and the textual units rated as most important. The analysis of the RS requires a rhetorical parsing algorithm based on a corpus analysis of discourse markers and text fragments. The rhetorical analyser locates the most salient textual units and structures.

Frames/templates. The design and production of frames is a technique of natural language processing for the extraction of information, whenever the information is structured in recognizable patterns within a digital setting [32]. Systems such as SCISOR [33] or JASPER have been developed [34] for the extraction of news items in the financial press by means of abstract frames in combination with techniques of partial analysis. FIES extracts Financial Information from Electronic Sources [35]. Frames/templates may also be used in summarizing scientific documents [36], information extraction from chemistry documents [37] and automatic bibliographic reference extraction from full-text patents in English [38].
Some progress in the processes of information reduction came about with the development of the ADAM system, centred on chemical information analysis [39]. As chemical texts tend to be highly structured and the semantic scheme is quite predictable, abstract frames are used to identify the important concepts and produce indicative abstracts. These kinds of frames resemble contextual models in the way they help identify concepts – bearing in mind the elements of the rhetorical structure – and the associated semantic roles. The program registers rhetorical indicators such as 'the results indicate', so that these sentences are automatically included in the abstract. When anaphoric-type references cause problems of semantic coherence, the ADAM system can resort to a WCL. This WCL contains some 700 cues – each associated with a syntactic value and a semantic weight – and this information is used to decide whether or not to include a sentence in the abstract.

The multi-lingual PROTEUS system (PROtotype Text Understanding System) was designed to analyse and interpret journalistic texts in English and Spanish language pairs, and produces extracts in a structured form [40]. Like other examples based on NLP, it operates with three components: a syntactic analyser,

with parser, lexical analyser and compiler; a semantic analyser; and an abstract frame generator. PROTEUS carries out the syntactic analysis of each sentence by generating a conceptual structure in which each term is related to a concept, with the different concepts organized into a hierarchical structure. This syntactic structure is then processed by the semantic analyser, which assigns each term to a thematic group and tries to fit the sentence structure into one of the semantic models. Finally, the abstract frame generator takes the semantic information from the system and transfers it to the database fields.

The use of abstract frames has become commonplace on the Web as a tool for processing electronic information and its subsequent extraction or retrieval. Altavista, for instance, often uses frames with natural language sentences related to the search topic, so that the user can select the most appropriate option. The metasearcher *Ask Jeeves*, based on techniques of knowledge management, gathers the experience of expert human searchers on the Web, and organizes this information in a database. Professional editors review Web resources in order to build a knowledge base about sites that might be consulted for common queries, and the resulting lists of questions with the corresponding Web pages for answers are stored in the knowledge base by means of the abstract frames.

Self-formatting procedures are one way of dealing with the proliferation of electronic documents while avoiding the high labour cost for their processing. Groups of metadata to be included in the source document should indicate the key concepts and their categories by means of standardized automatic frames.

4.2. Cognitive summaries

From a cognitive standpoint, taking semantic models into account carries out the processing of contents and subsequent generation of summaries. Most of these summarizing systems are based on the Kintsch/Van Dijk [41] hypotheses and strategies. Humans understand a text by interpreting and reconstructing its meaning according to their base of knowledge. Inference, the root of human comprehension, takes advantage of that previous knowledge, adapting it to the new information and filling the gaps of coherence. The interpretation is represented in memory as the basis for possible summaries. Among a varied methodology, some examples deserve special mention.

The SUMMONS system (Summarizing Online News articles) is centred on information about an event from different informative sources. Sets of templates are introduced in the content planner, which selects the information to be included in the summary. Different operators are then activated [42] as the change of perspective, contradiction, addition, refinement, agreement, superset and no information. Finally, a linguistic generator determines the words and the syntactic form of the summary. The NLP group at Columbia University has developed other systems of summarizing along these lines [43] such as MultiGen, for domain-independent multiple documents (1999), FociSum, for documents separated from a domain by focal analysis (1998), and SumGen, a 'cut and paste' method for extracted sentences (1999).

FRUMP [44] is a program for summarizing short articles which mirrors the human processes of interpreting new events from the data and expectations held in one's personal cognitive frame. Within FRUMP, expectation leads interpretation, and a situational knowledge base is used to predict general events from a text analyser.

The SUSY [45] program is centred on specialized scientific text. Target summary frames are retrieved and adapted from files of basic texts and summary structures, and a parser/syntactic analyser acts on three levels. At the stage of sentence understanding, a propositional representation is constructed. For the analysis of the text structure, that basic linear representation is extended within a broadened linear representation. Finally, the elements of the extended linear representation are ordered, giving a hierarchical propositional network that reflects a deeper knowledge of the source text. The summary is further elaborated with reference to sentence models that follow the basic rules of abstracting style.

Over the last two decades, some systems for the analysis of short texts in very restricted domains, which are based on artificial intelligence, have been developed. The first noteworthy achievement of this sort was SCISOR (System for Conceptual Information Summarization Organization and Retrieval) conceived by Rau [46]. This prototype operates in the management domain, processing news from online information sources, and items published in periodicals such as the *Wall Street Journal*. The documentary information as well as that from the questions/user queries is stored in Kodiak language. Unlike conventional retrieval systems, SCISOR allows for the retrieval of conceptual information, answering simple questions formulated in natural language. The system prerequisites consist of a very broad knowledge base that includes the domain vocabulary and a phrasal lexicon. The architecture

of SCISOR is quite comprehensive and follows this sequence:

- The selection and filtering of news in natural language obtained from online documents, interactive tools and techniques of automatic acquisition.
- The syntactic analyser operates on two levels – the full or comprehensive grammatical analysis, for precise semantic interpretations; and the partial or superficial analysis, focusing on the extraction of words taken from the text. The full analysis is performed by TRUMP, a portable package of comprehension programs that includes a syntactic analyser and a semantic interpreter. The partial analysis is carried out by the TRUMPET subsystem.
- Finally, the application of KING (Knowledge Intensive Generator) creates different-sized summaries made up of a categorical selection of natural language sentences. The mapping entails the retrieval and application of structured associations that relate the concept to be expressed with other conceptual and linguistic structures. The selection of templates/frames allows the selected grammatical structures from the knowledge base to be combined with the conceptual structures obtained from the mapping.

KING and TRUMP were designed and developed to provide an integrated system of language processing. Although SCISOR does not produce authentic abstracts, this experimental prototype has brought some progress in the comprehension, representation and subsequent retrieval of short texts from restricted knowledge domains.

Endres-Niggemeyer, also working from a cognitive perspective [47], proposes a naturalistic model of automatic abstracting based on the combination of the experience of professional abstractors and the KADS methodology (expert model-driven knowledge engineering). Working upon the well-known propositions of the sector (Fidel, Creemins, Lancaster, Pinto, etc.), and carefully analysing the basic steps in expert abstracting, the model emulates human abstracting operations. The system is known as SIM-SUM and describes the set of automatic tools based on artificial intelligence for the elaboration of summaries in restricted domains. Based on a blackboard cognitive architecture, it allows the integration of very local and centralized activities resulting from the complex interaction of several simple modules. The Rhetorical Structure Theory supports SIM-SUM, as it offers structural clues as to the conceptual and organizational scheme of the different text units, assigning them to a sequence [48]. Since abstractors face ever-increasing amounts of information with less and less time to carry out their task, the exploratory techniques they use tend to be partial, and top-down. A first and rapid reading is made to get a general idea of the higher level of the rhetorical structure, as in the introduction, which usually offers a representation of the semantic text structure. From there, the abstractor would move on to the body of the document, with its methodology, results and conclusions, which may be clearly discernible in the text. The SIM-SUM system, then, facilitates documentary content modulation and summary production, albeit in very specific domains.

Skoroxod'ko [49] conceived an automated abstracting method based on the adaptation to the text source. All texts have some individual characteristics of their own, reflected in semantic networks representative of the content. Hahn and Reimer [50] describe the procedure of an expert system for condensing documents that relies on a basic knowledge network applied to the text.

Summaries are not only an abbreviated documentary form but also a tool for assessing the relevance of a given document to a selected topic, reducing the amount of information to be read. However, even assuming that the method of identifying important content has worked correctly, the final output summary is best described as a semi-text indicative summary which notes the main topics of the source document. It is clear that some method for organizing and ordering the output material is needed [9].

5. Graphic and relational methods

The semantic network (concept mapping) of the source document is also an appropriate mode of document representation. One of the most important developments in this field is the SMART system, a tool for information retrieval that uses vectorial spaces [51]. Vectors, or groups of weighted terms, represent all the informational units (text and queries). The terms may be selected from a controlled vocabulary list, from a thesaurus, or directly from the text or query in question. The well-known formula $(f_t \times 1/f_c)$ favours terms with a high frequency in documents (f_t), but with a low frequency in the collection (f_c). Once texts and questions have been represented by means of vectors of weighted terms, a measure of similarity between pairs of vectors can be computerized, from 0 for unlinked to 1 for identical vectors.

Texts are heterogeneous not only by their content, but also by length. In a vectorial environment, it will sometimes be difficult to find a vector of similarity between pairs of texts, because of a significant difference in their extension. Processing lengthy documents is awkward, and breaking them down into passages can prove advantageous. The SMART system allows the re-structuring of complete texts or fragments of variable length (sections, paragraphs, groups of adjacent phrases or loose sentences). Relationships between passages of texts (with similarities that score over a pre-established value) can be expressed in relational maps. The closeness of the information nodes on the map would reflect the basic contents of a text.

This model of automatic text analysis entails the phases of text theme identification, selective crossing, and relevant parts extraction:

- Identification of the textual topic – topic or text (related texts) subject matter is the first step in approaching content. Although some linguistic-cognitive methods allow the theme to be identified, the vectorial model uses the relational maps as elements of entry, to cluster together groups of text passages that are closely interrelated yet relatively disconnected from the rest. To identify the theme, triangles (group of three related text passages) of the relational map are discerned and a central vector is drawn for each triangle.
- Text selective crossing – as in a network, one of three inner routes may be taken through the text, depending on the user's priorities: the weight of the nodes that have a greater number of links; the detection of text passages in strategic positions, such as first paragraphs of each section, or the first sentences of each paragraph; and the weight of the links.
- Extraction of text relevant parts – applying ideas from hypertext, intra-links can be created between the text paragraphs or sentences. By locating these paragraphs and links on a relational map, the structure of the document can be visualized and consequently discovered by means of the appropriate automatic techniques. The best summaries are generated by extracting the most central passages, that is, those with a greater number of semantic relations [52]. One important structural property of most texts is the use of multiple functional units. To automatically locate such units or segments – contiguous pieces of text that are well linked internally whilst very disconnected from the adjacent text – it is necessary to find gaps between adjacent paragraphs on a relational map. When a passage is well-linked internally but is not substantially linked to the surrounding text, the passage probably deals with a coherent topic within the article and constitutes a functional unit. The segments or nodes with many links to other nodes are known as bushy, and the extract may be elaborated by selecting those bushy nodes and maintaining the order in which they appear in the text. However, the results obtained might not satisfy all users, given the variety of their needs and the different parts of the source document that might be considered most relevant for the summary.

A problem in most databases is the multiplicity of documents with a common theme. In the case of related documents within a collection, Mani [53] proposes a method of representation based on the finding similarities and differences (FSD) algorithm to pairs of documents. The concepts of words, proper names and phrases are represented by location, as nodes in a graph. Corresponding links set semantic and topological relationships among conceptual units. If we are thus able to discover, for a given theme and pair of related documents, all the nodes that are semantically related with the selected theme, then these nodes and their relationships can be compared to establish similarities and differences between the documents. In this way, once the common themes from the intersection of the activated concepts are selected within each graph, summaries of several documents can be produced. This technique may be particularly useful when applied to an unlimited number of texts, as might be the case with Web pages.

In the light of these findings, we can say that even the most sophisticated automatic methods are far from satisfactory in terms of time, cost of processing and results obtained. Even with a relatively small corpus (some 1500 messages) of short texts (around 14 sentences), the best of methods will encounter difficulties in producing optimal results, as not all the relevant sentences are selected, nor all the selected sentences are truly relevant [54].

6. Hybrid methods of production: pre-abstract and 'customized' abstracts

In view of the observed deficiencies of the many automatic abstracting systems, it seems unlikely that any single autonomous program would be widely

accepted. It is much more reasonable to imagine the human abstractor taking part in the process. Thus, our search for an efficient process of production and an effective product leads us to a machine-assisted abstract system specially designed to serve as a workbench for the human abstractor [55]. Upon entering a new source document, the system should automatically compose and display an initial pre-abstract, pointing out all the possible points of interest detected in the source document, inserting links when it detects strong connections between separate parts, underlining any anaphora and displaying a list of possible antecedents. The abstractor can re-order, erase or edit the sentences in the pre-abstract, add new ones, consult any relevant index or thesaurus, and save the present state of the pre-abstract. So, there is the possibility of preparing two or more different (customized) summaries depending on their extension and scope.

TEXNET is an experimental hybrid system to help the human abstractor, responsible for certain tasks while the computer program takes care of others [56]. Among the pending research fronts for this system are tools for parsing and automatic recognition of relevant characteristics of source texts, its structured exposition and conceptual networks derived from them. There is also the question of integration of the thesaurus and the production of extracts and other intermediate forms. Experiments were carried out to evaluate the general usefulness for the abstractor, to formulate preliminary hypotheses about abstractor–software relationships and reactions, and to come up with ideas for further developments. The abstractor workbench is a promising idea, and computer programs such as TEXNET will surely facilitate the application of all the abstracting research to unfold in the near future.

7. Conclusions

There is much divergence between the different approaches to the problem of the automation of abstracts. For this reason, we believe that a first step toward the improvement of the automatic methods would be some theoretical consideration that would perhaps help to put a rather confused panorama into some kind of order.

Abstracting is a cognitive and pragmatic activity that is highly conditioned by the contexts, and the use of abstracts is also context-dependent. As a result, there is a need for a prior definition of the contexts of use and production of abstract and abstracting: establishing the type of user, the available means, the type of document source, the abstracting goals and the type of derived product. Without a definition of these points, it will not be possible to establish an efficient and reliable methodology. The abstract's level of description is related to the context of use, and different levels of granularity produce various abstracts, according to the open and plural document that we are looking for in order adapt it to the various user needs. A source document may generate as many abstracts as the number of different contexts of use and production.

The context of cognition is also relevant: we need to recognize that each concept domain has a content of it own with an individual ontology and a characteristic rhetoric. In this respect the evolution is towards a non-static and non-lineal structure. The abstract of the future may be a reticular and/or graphic product that could ease the visualization and the immediate access to information. The engineering of abstracting would increase the character of abstract as tool of navigation to the detriment of its informative capacity. In any case, abstracts, in view of their informative richness, are destined to be among the richest metadata of the future and perhaps the resource discovery metadata par excellence.

However, this contextual knowledge is not enough: even with the richest knowledge of both kinds of contexts (use-production and cognition), we have to recognize the human presence as the only way to provide the abstract with the necessary coherence and intention. Instead of being merely a machine-produced metadata, the abstract must be elaborated by humans, with the collaboration of computers. This is because machines can handle data and even information, but knowledge is an exclusively human activity (Fig. 1).

The full automation of abstracts has been dismissed as an unrealistic dream, thus we have to recognize the effectiveness of other forms of auto-representation, such as extracts and summaries. However, the representativity of these, although effective in some situations, is limited and less than that of the abstracts. Consequently, the research into abstracts continues to be worthwhile, as the more complete form of document representation. Fortunately, a partial automation system based on man–machine collaboration is possible. The role of the abstractor, whose category is applicable to that of the author, is unavoidable. Given his or her importance, the documentary support of this activity should be given special care. The introduction of new technologies – namely computers and telecommunications – has altered the scenario of the professional abstractor who can now use these innovations to

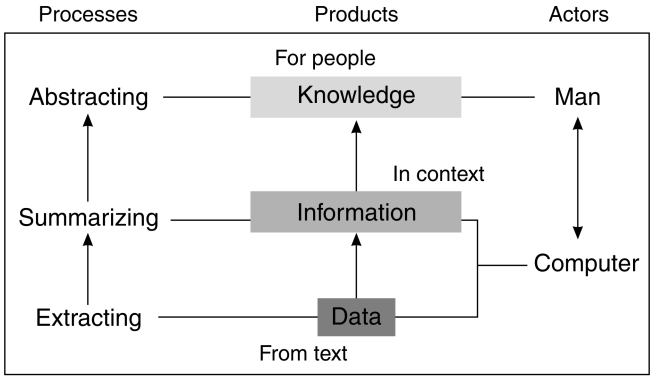

Fig. 1.

improve productivity. The presence of personal computers and automated extracting and summarizing techniques are the best-known contributions.

A thorough process of auto-abstracting planning would need to keep the following ideas in mind:

- In the study of automatic representation of textual information, and specifically that of the automatic abstract, it is helpful to highlight the following levels, on ascending order of difficulty and depending on the perspective taken: morpho-syntactical (physical), pragmatic (operative) and semantic (cognitive).
- From a physical perspective, the possibility of automatically representing information is optimal, because manipulated material entities have a maximum level of definition. However, the reduction of physical objects by means of extraction may be inadequate with respect to the attainment of knowledge.
- Assuming the incapacity of the programs to manage efficiently the isolated data from a cognitive standpoint, we have to recognize the need for work with organized and contextualized data.
- Yet this is not enough. In order to produce authentic abstracts, human activity is essential and the role of the abstractor is, at least for the time being, unavoidable, because of the interaction between the aforementioned cognitive and pragmatic levels of information processing.

References

[1] S.L. Vellucci, Metadata, *Annual Review of Information Science and Technology (ARIST)* 33 (1998) 187–222.
[2] A. Chilvers and J. Feather, The management of digital data: a metadata approach, *The Electronic Library* 16(6) (1998) 365–371.
[3] R. Heery, A. Powell and M. Day, *Metadata, Library & Information Briefings*, Vol. 75 (South Bank University, London, 1997).
[4] H.P. Luhn, The automatic creation of literature abstracts, *IBM Journal Research and Development* 2(2) (1958) 159–165.
[5] B. Mathis and J. Rush, Abstracting. In: E. Dym (ed.), *Subject and Information Analysis* (Marcel Dekker, New York, 1985).
[6] E. Liddy, Natural language processing. In: P. Atherton and E. H. Johnson (eds), *Visualizing Subject Acces for 21st Century Information Resources* (University of Illinois, Urbana-Champaign, IL, 1998).
[7] J. Allen, *Natural Language Understanding* (Benjamin Cumming, 1995).
[8] M.F. Moens, *Automatic Indexing and Abstracting of Document Texts* (Kluwer Academic, Boston, MA, 2000).
[9] K. Sparck Jones, Automatic summarising: factors and directions. In: I. Mani and M. Maybury (eds), *Advances in Automatic Text Summarization* (MIT Press, Cambridge, MA, 1998).
[10] G. Salton, *Automatic Text Processing* (Addison-Wesley, New York, 1989).
[11] C.J. Armstrong and A. Wheatley, *A Survey of the Content and Characteristics of Electronic Abstracts* (Library Information Technology Centre, London, 1997).
[12] A. Gilchrist, Who is responsible for information quality in the information society? In: *5th Jornades Catalanes de Documentacio* (Barcelona, 1995).
[13] J.L. Milstead, Thesauri in a full text world. In: P. Atherton and E.H. Johnson (eds), *Visualizing Subject Access for 21st Century Information Resources* (University of Illinois, Urbana-Champaign, 1998).
[14] J. Qin and K. Wesley, Web indexing with meta fields: a survey of web objects in polymer chemistry, *Information Technology and Libraries* 7(3) (1998) 149–156.
[15] L. Dempsey and R. Heery, Metadata: a current view of practice and issues, *Journal of Documentation* 54(2) (1998) 145–172.
[16] A. Wheatley and C.J. Armstrong, Metadata, recall, and abstracts: can abstracts ever be reliable indicators of document value? *Aslib Proceedings* 49(8) (1997) 206–213.
[17] J.G. Kircz, Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval, *Journal of Documentation* 47(4) (1991) 354–372.
[18] H. Borko and Ch. Bernier, *Abstracting Concepts and Methods* (Academic Press, New York, 1975).
[19] C. Paice, Constructing literature abstracts by computer: techniques and prospects, *Information Processing and Management* 26(1) (1990) 171–186.
[20] J. Kupiec, J. Pedersen and F. Chen, A trainable document summarizer. In: E.A. Fox, P. Ingwersen and

R. Fidel (eds), *Proceedings of the 18th ACM-SIGIR Conference on Research and Development in Information Retrieval* (Seattle, WA, 1995).
[21] S. Teufel and M. Moens, Sentence extraction as a classification task. In: I. Mani and M. Maybury (eds), *Intelligent Scalable Text Summarization. Proceedings of a Workshop Sponsored by the Association for Computational Linguistics* (Madrid, 1997).
[22] IBM Corporation, *Advanced Systems Development*, ACSI-Matic Auto-Abstracting Project, Final report 1–3 (1960–1961).
[23] H.P. Edmundson, New methods in automatic extracting, *Journal of the Association of Computing Machinery* 16(2) (1969) 264–285.
[24] H.P. Edmundson, V.A. Oswald and R. E. Wyllys, *Automatic Indexing and Abstracting of the Contents of Documents* (Planning Research Corporation, Los Angeles, CA, 1959).
[25] J.E. Rush, R. Salvador and A. Zamora, Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria, *Journal of the American Society for Information Science* 22(4) (1971) 260–274.
[26] R. Brandow, K. Mitze and L. Rau, Automatic condensation of electronic publications by sentence selection, *Information Processing and Management* 31(5) (1995) 675–685.
[27] R. Barzilay and M. Elhadad, Using lexical chains for text summarization. In: I. Mani and M. Maybury (eds), *Intelligent Scalable Text Summarization. Proceedings of a Workshop Sponsored by the Association for Computational Linguistics* (Madrid, 1997).
[28] H. Ahonen, Knowledge discovery in documents by extracting frequent word sequences, *Library Trends* 48(1) (1999) 160–181.
[29] B.A. Mathis, Techniques for the evaluation and improvement of computer-produced abstracts. Technical Report OSU-CISRC-TR-72 (Ohio State University, Columbus, 1972).
[30] K. Ono, K. Sumita and S. Miike, Abstract generation based on rhetorical structure extraction. In: *COLING 94, Proceedings 15th International Conference on Computational Linguistics* (Kyoto, 1994).
[31] D. Marcu, From discourse structures to text summaries. In: I. Mani and M. Maybury (eds), *Intelligent Scalable Text Summarization, Proceedings of a Workshop sponsored by the Association for Computational Linguistics* (Madrid, 1997).
[32] G. Chowdhury, Template mining for information extraction form digital documents, *Library Trends* 48(1) (1999) 182–208.
[33] P. Jacobs and L.F. Rau, SCISOR: extracting information from on-line news, *Communications of the ACM* 33(11) (1990) 88–97.
[34] P.M. Andersen *et al.*, Automatic extraction of facts from press releases to generate news stories. In: *Third Conference on Applied Natural Language Processing* (Trento, 1992).
[35] Chong, A.G., FIES: financial information extraction systems, *Information Services and Use*, 17(4) (1997) 215–223.
[36] P.A. Jones, C.D. Paice, A 'select and generate' approach to automatic abstracting. In: T. McEnery and C.D.Paice (eds), *Proceedings of the BCS 14th Information Retrieval Colloquium* (Springer, Berlin 1992).
[37] G.J. Postma and G. Kateman, A systematic representation of analytical chemical actions, *Journal of Chemical Information and Computer Sciences* 33(3) (1993) 350–368.
[38] N. Lawson, M.F. Kemp, G.G. Lynch and Chowdhury, Automatic extraction of citations from the text of English language patents: an example of template mining, *Journal of Information Science* 22(6) (1996) 423–436.
[39] C.D. Paice and P.A. Jones, The identification of important concepts in highly structured technical papers. In: *SIGIR-93: Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Association for Computing Machinery, New York, 1993).
[40] R. Grishman, *Information Extraction from Natural Language Text*. PROTEUS Project Memorandum, no. 47. (Department of Computer Science, New York, 1991).
[41] T.A. Van Dijk and W. Kintsch, *Strategies of Discourse Comprehension* (Academic Press, New York, 1984).
[42] K. Mackeown and D.R. Radev, Generating summaries of multiple news articles. In: E.A. Fox, P. Ingwersen and R. Fidel (eds), *Proceedings 18th ACM-SIGIR Conference on Research and Development in Information Retrieval* (Seattle, WA, 1975)
[43] Columbia Natural Language Projects; www.cs.columbia.edu/nlp/projects.html (access date July 2002).
[44] G. Dejong, An overview of the FRUMP systems. In: W.G. Rehnert and M. Ringle (eds), *Strategies for Natural Language Processing* (Lawrence Erlbaum, London, 1982).
[45] G. Fum, C.A. Guida and A. Tasso, A prepositional language for text representation. In: B.G. Bara and G. Guida (eds), *Computational Models of Natural Language Processing* (North-Holland, Amsterdam, 1984).
[46] L.F. Rau, Knowledge organization and access in a conceptual information system, *Information Processing and Management* (23)4 (1987) 269–283.
[47] B. Endres-Niggemeyer, *Summarizing Information* (Springer, Berlin, 1998).
[48] W.C. Mann and S.A. Thompson, *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87–190 (1990).
[49] E.F. Skoroxod'ko, *Adaptative Method of Automatic Abstracting and Indexing* (IFIP Congress, Yugoslavia, 1971).

[50] U. Hahn and U. Reimer, Heuristic text parsing in 'TOPIC': methodological issues in a knowledge-based text condensation system. In: *Representation and Exchange of Knowledge as a basis of Information Processes* (North-Holland, Amsterdam, 1984).
[51] G. Salton, J. Allan, C. Buckley and A. Singhal, Automatic analysis them generation and summarisation of machine-readable texts. In K. Sparck Jones and P. Willet (eds), *Reading in Information Retrieval* (Morgan Kauffman, San Francisco, CA, 1997).
[52] G. Salton A. Singhal, M. Mitra and C. Buckley, Automatic text structuring and summarization, *Information Processing and Management* (33)2 (1997) 193–208
[53] Mani and E. Bloerdorn, Multi-document summarization by graph search and matching. In: *Proceedings of American Association for Artificial Intelligence* (1997), pp. 622–628.
[54] F.W. Lancaster, *Indexing and Abstracting in Theory and Practice* (Graduate School of Library and Information Science, University of Illinois, Champaign, IL, 1998), p. 293.
[55] C.D. Paice, Automatic abstracting, In: *Encyclopaedia of Library and Information Science* (Marcel Dekker, New York, 1994), Vol. 53(16), pp. 16–27.
[56] T. C. Craven, Abstracts produced using computer assistance, *Journal of the American Society for Information Science* 51(8) (2000) 745–756.