ISKO: La Representacion y Organizacion del Conocimiento: Metodologias, Modelos y Aplicaciones, Madrid, 25-27 April 2001, p. 32-41

COGNITIVE PERSPECTIVES OF REPRESENTATION

Autores: Peter Ingwersen
Royal School of Library and Information Science
pi@iva.dk

Resumen: El trabajo introduce la concepción cognitiva de poli-representación o multievidencia aplicada a la recuperación de información, en particular asociada con la representación documental. Se describen y discuten varios tipos de conocimiento: el del autor, el conocimiento del indizador así como otros tipos de características documentales de naturaleza representativa. Se discute también la presunción de que en
la relevancia de la recuperación inciden aspectos cognitivos y funcionales o pragmáticos y se analiza la utilidad del clustering para representaciones complejas con funciones de navegación o visualización.

Palabras Claves:
Representación del conocimiento; recuperación de información: modelos cognitivos; pragmática.

Abstract: The paper introduces the cognitive conception of poly-representation or multi-evidence applied to information retrieval, in particular associated with representation of information objects. Various types
of aboutness are described and discussed, i.e. author and indexer aboutness, as well as other forms of document features of representative nature. The assumption that highly relevant objects are found in the retrieval overlaps of cognitively and functionally different origin is discussed and the utility of clustering of objects by complex representations for navigation or visualisation purposes is briefly analysed.

Key Words: Knowledge representation; information retrieval; cognitive models; pragmatic.

1. Introduction

Knowledge organisation, information retrieval and informetrics are interwoven sub-disciplines of information science. Obviously, information retrieval is necessary for informetric, bibliometric or scientometric analyses for data collection purposes (Christensen & Ingwersen, 1996). Information retrieval relies in many ways on bibliometric laws, such as Zipff´s Law on term frequency in text corpora, and share clustering models, such as the Vector Space Model and the use of similarity measures (Salton & McGill, 1983). For both subdisciplines knowledge or document representations are crucial for success. The aim of this paper is to point to the *cognitive variety* of representations in information objects that are useful for retrieval purposes in full-text[1] digital environments by improving the intellectual access to knowledge sources.

[1] Throughout the paper full-text signifies all kinds of "full" information objects, e.g. images, video, etc.

The idea of exploring the potential value of matching the multidimensional cognitive variety of representations inherently existing, extracted or interpreted from information objects *and* from the cognitive space of a user originates from Ingwersen (1996). In his cognitive theory for interactive information retrieval the notion of poly-representation or multi-evidence is introduced. Poly-representation is defined as a variety of different pre-suppositions and interpretations of situations made by the different cognitive agents that take part in the processes of information transfer. In information retrieval such agents are predominantly authors, indexers, algorithmic or computational designers, thesaurus
constructors, interface designers, journal and database editors, and users. Each agent contributes his or her cognitive interpretation of the information situation, in a social context, to the representations of available information objects (or documents). Hence, the representations are of different cognitive origin, also over time, or from the same origin but of different functional nature, for instance, author generated text, diagram captions and references or out-links. The representations can be made in different presentation styles according to domain and media. Following the cognitive theory of information science and retrieval the processes of information transfer are seen as processes of cognition in which the variety of representations acts as supplementary contexts to one another. A second notion becomes thus important, i.e. that of *cognitive overlaps* of different representations. The paper seeks to extend the model of such overlaps.

Cognitive overlaps, produced during processes of information retrieval, imply that representations of different cognitive origin or different functional nature point to the same information objects. The more different in cognitive origin the higher the probability that such objects are *relevant*. This assumption is based on very few experiments carried out in the citation analysis research environment, for instance, by Pao (1994). In her investigation Pao intersected sets of bibliographic records retrieved by index and title terms with sets retrieved by citation analysis based on an initial pertinent seed document. The intersection, i.e., in a cognitive sense the document overlap made of different cognitive origin from authors, indexers, and citing authors, was then evaluated by domain experts for topical relevance - but without the experts knowing from which sets the documents derived. Pao found that the density of relevant documents in the overlap was more then six time higher than in the original separate sets. We are currently initiating tests to explore further the application of cognitive retrieval overlaps in full-text environments.

In a cognitive theory for information retrieval the *user* and his or her representations of information need situation, current cognitive state and interpretations of the work tasks or interest underlying the information need situation in a social or organisational context are central elements. In the cognitive space of the user the request formulation is a representation of the user´s current cognitive state concerned with an information need. Similarly, conceivable problem statements and work task or interest descriptions are representations of intrinsic cognitive structures underlying the user´s information seeking. Such intrinsic cognitive structures of users are more or less variable and transformable, as opposed to system and text representations that, in a cognitive sense, always are invariable when stored at a given point in time. From this perspective information retrieval sessions are frequently exploratory throughout long initial phases (Bates, 1989). Real-life information needs may thus be variable, initially vaguely stated or ill-defined. Well-defined and static wishes for information seem to be a special case (Schamber et al., 1990). Structures of knowledge representations, and most probably also the visual presentation of such representations and structures, are hence of central importance for the cognitive support of the user during retrieval. Indeed, topical retrieval is not the only way to reach into the information space. Many other access points and modes of representation are often available which, together with subject matter, may form part of the user´s knowledge and wishes for information. However, the remaining of the paper will concentrate on the phenomena of representations of information space and not pursue further the cognitive space of the user.

The paper is structured as follows. The next section explores the types of aboutness leading to the variety of cognitively different types of representations concerned with subject matter as well as other facets of information objects, illustrated by representation samples from scientific communication. This is followed by a discussion of the association between variation of representation and relevance conceptions as well as a brief illustration of alternative or additional ways of representing and visualising information objects, and concluding remarks.

2. Types of aboutness and isness

In (Ingwersen, 1992) a typology of aboutness is proposed which associates to the original meaning of aboutness put forward by Hutchins (1978). The typology operates with the ollowing categories:

- Author aboutness, i.e. the *contents as is*;
- Indexer abouness, i.e. the *interpretation of contents* with a purpose;
- Request aboutness, i.e. the user or intermediary interpretation or *understanding of the information need* as represented by the request.

Author aboutness signifies the contents of the information objects in the form of signs, i.e., the transformations of the interpretations, ideas, and cognitive structures of the author(s) with respect to their goals and intentionality. If we consider scientific communication by means of articles or monographs, the contents (and signs) is text, commonly structured in specific ways according to convention, e.g. introduction, theory or methodological sections, results, and discussion or conclusions. In addition, we have some references to related work that have been interpreted in some way by the author(s). Like for diagrams or figures, and their captions, references signify cognitively different ways of representing the object. The application of references given to other related work as an *alternative form* of knowledge representation, which can be used for automated retrieval purposes was originally invented by Garfield in the 60s in connection with his citation index approach (1998). We should observe that what we are counting operationally when performing citation analysis is not references but citations received by a scholarly entity, like an author, an article, an institution or a country. This distinction is important because commonly information objects in scholarly communication list a fixed number of references (or changeable number of out-links on the Web) to not only fellow scientists but also to particular articles and journals carrying those articles. Over time, however, the same object may become focus of attention by being increasingly cited (or linked to) by *other authors*. We may thus talk of two different but inter mingled networks of references and citations. Their cognitive and functional nature is quite different from one another and from that of the text itself.

Author aboutness consequently provides the full text, the document structure, the chapter and section titles, the captions, and the references as means for automated indexing. In modern (topical) information retrieval the words from the text corpus are weighted according to a scheme (Van Rijsbergen, 1979). However, weighting of the references has *not* been attempted due to the online citation database structures and their lack of full-text. All references weight the same. In full-text scholarly information systems, like those of the large publishing houses or ResearchIndex.com, references or authors referred to could easily be detected, counted for frequency and weighted – like single words or noun phrases in the same text.

Author aboutness and the contents-as-is representation provide many and different access points to information objects. The vocabulary is that of the domain as interpreted by the author(s). Supposedly, that vocabulary corresponds to that of future users. However, this is no guarantee that the *semantic* contents correspond to the users´ information needs. During retrieval all the original intentionality of the author(s) as well as the meaning of the text is fundamentally *lost*. It can only be recovered by the interpretation of users. In order to ease meaningful access to the information objects alternative or supplementary human representations have been introduced in the form of classification and indexing, i.e., human indexer aboutness. A recent article looks specifically into the nature of indexing both from a human perspective and in an automatic sense (Anderson & Pérez-Carballo, 2001.

Indexer aboutness is based on human interpretation of information objects, different from that of authors. Aside from topically classifying objects, indexers may add new perspectives to the contents of such objects and ease the access to meaning. Cognitively speaking – and in practice – human indexing is directed towards the entire document. This is logical when categorizing the document type or information mode. However, the actual number of indexing terms or phrases added to the contents description of the information object is commonly very limited. This leads to *reductionism*, in particular when alone the major themes and aspects are represented by the indexing. Another drawback is the difficulty of applying weighting schemes to indexing structures. Too few term occurrences exist for adequate weighting purposes. Inter-indexer-inconsistency – also over time – is also an issue. However, in a cognitive sense inconsistency is preferable to consistency due to a wider range of available access possibilities (Ingwersen, 1996). Indexing may also be supportive during retrieval by cleaning up the name form mess increasingly experienced in databases and knowledge resources.

Indexer aboutness can be directed towards the subject matter and meaning of the information object or towards its future potential intellectual use or user grouping(s). The former purpose is the common issue of indexing while the latter calls for tremendous predictive power of the domain expert indexer. Some information researchers regard the socio-scientific domain as the determining force during the indexing process (Hjorland, 1997; Jacob & Shaw, 1998). From a cognitive perspective the indexer interprets the information object situated in the domain(s) and *influenced* by the social and scientific (historical) context. However, the indexer determines the interpretation – not the social scientific construct surrounding the object and indexer at a given point in time. A recent interesting analysis by Jacob discusses two approaches to classification in situated context (2001). Classification and similar knowledge organizational structures are seen as functionally different from the process of indexing.

Designing classification systems or *generating thesaurus structures* are processes that are cognitively different from the processes of classifying and indexing, for instance, using a structured controlled vocabulary from a domain thesaurus. Such structures are socio-cognitive by nature, as they are negotiated at a certain point in time by a team of domain experts. As such, a domain thesaurus displays *cognitive authority* – at least for while. In an authoritative way, thesauri and domain specific author co-citation maps are similar. Both derive from interpretations of the domain in question, but author co-citations are dynamic and the document clusters are changing over time. Many other forms of representations might also be applied to clustering of a domain, for example, term co-occurrence in full-text documents. Figure 2 below displays a map that clusters documents according to *national research profiles* each consisting of the publication counts from nine different sub-disciplines of the social sciences.

Figure 1 demonstrates the variety of cognitive aboutness structures, dealing with aspects of subject matter, *and* structures of different interpretative or assessing nature, i.e. they are *selective* in socio-cognitive ways different from those of indexers. Instead of aboutness they reflect *isness* by making available non-topical features inherent in information objects – depending on media, domain, and presentation style. In scientific communication articles or conference papers are commonly submitted for peer reviewing to be accepted or rejected by a (digital) journal or conference. The reviewing process is sociocognitive by nature, taking into account the paper and its scientific contribution, the journal or conference reputation and scope in context of the domain(s) treated by the journal or conference. First after a positive reviewing process does the journal or conference *select* the submission for publishing, i.e. the cognitive authority of the *journal name* becomes assigned the paper. How and when it is published, e.g. categorized together with other selected papers, is determined by the editors of the journal or the programme committee or chairs of the conference.

Also the author affiliation (and country) belongs to the isness features, together with database names later to index and incorporate the journal or conference as part of their structures. It is not uncommon that a particular topic is covered by hundreds of databases and that single articles can be found in several different systems due to the inclusion of the journal in the files.

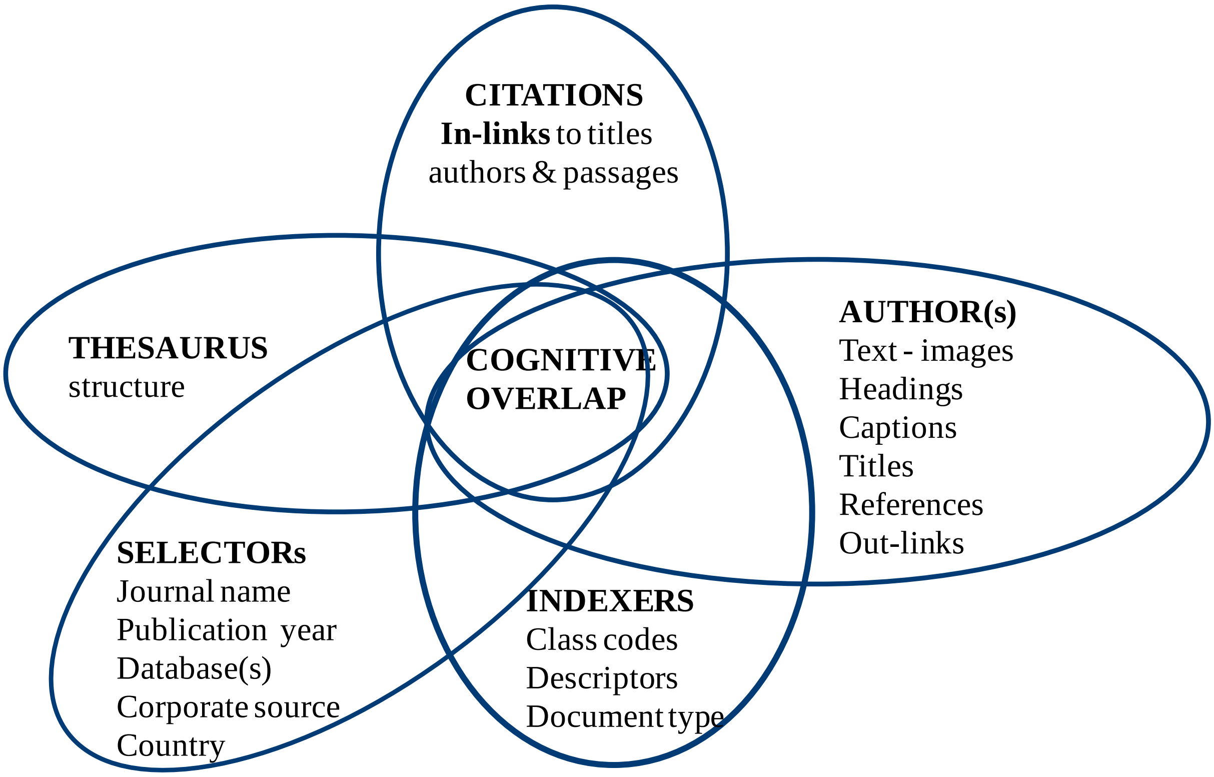

Figure 1. Poly-representative overlaps of cognitively and typologically different representations of information objects. Retrieved sets generated by one search engine and associated with one searcher statement. Extension of (Ingwersen, 1996, p. 28).

Aside from illustrating the different representations (or access points) associated with scholarly papers and described above, the figure points to the central cognitive as well as the many possible overlaps between the sets of information objects retrieved. The central cognitive overlap is defined by one retrieval engine that acts on one user statement, for instance, concerned with describing the work task situation underlying the information need. It extends the original version from 1996 by incorporating the additional features of isness related to the objects, such as journal name and publication date. Further, the logical separation of references (and out-links) from citations (and in-links) is carried out in the model.

From a cognitive stand, the objects found in the central cognitive overlap should be ranked prior to other objects since they are assumed most relevant. In principle, two or more retrieval engines can be fused, creating even more central overlaps. In addition, user statements causally associated with one another, e.g. the work task description and an information need request formulation, may form new overlaps.

We observe that the model takes into consideration that users may prefer to access retrieval systems partly by means of topical representations, partly via non-topical access points, like country or journal name. Since the central overlaps are considered highly relevant in accordance with Pao´s investigations (1994) the notions of relevance come into play (Cosijn and Ingwersen, 2000). Obviously, traditional topicality relevance assessments may take place in relation to the author and indexer aboutness representative structures. However, judgments of *pertinence* are also possible. Pertinence signifies the relation between the information objects and the user perception of the information need, for instance, in the form of novelty or cognitive authority of author, affiliation or journal. Indeed, references or out-links may signal a relevant understanding of the matter to the user. *Situational* relevance, meant as the relation between the work task situation as perceived by the user and the objects, i.e., their usefulness, is conceivable due to full-text availability including diagrams and figures. Lastly, *socio-cognitive* or contextual relevance assessments are feasible by means of the citations (or in-links) given to the objects. The citations by scholarly colleagues imply commonly a certain degree of recognition, acceptance, and cognitive authority.

It is thus important to observe that all information objects can be represented by all their features, added or inherent, depending on media, domain and style; but in addition, the features may be represented by the objects. This is useful for the purpose of mapping by means of multidimensional scaling (MDS) techniques. Domain maps serve as tools for relevance judgments and feedback to retrieval systems. Mapping may further be applied for navigation purposes or as instruments for visualisation of information spaces.

Figure 2 demonstrates a map of visualisation of 17 selected OECD countries clustered according to the similarity of their publication profiles in the social sciences, 1994-98. Data derives from the National Science Indicators database (on CD-ROM), produced by Institute for Scientific Information, and containing data from the citation databases. The profiles consist of nine fields, including economics, sociology, political science, management research, social work, and library and information science. The map is an atypical multirepresentatio of a domain. The common practice in informetrics is to apply author co-citation, term co-occurrence, or even journal co-citedness. However, the current MDS map is made from the publications classified into nine social science sub-fields *and* affiliated to their author countries. The cosine similarity measure is applied to a matrix of 17x17 vector representations. The map illustrates nicely the dominance of the Anglo-American publications in the domain, but also that some clusters of smaller EU countries begin to form to the East of the core cluster. The map signifies that countries located near each other have similar research profiles. In an informetric sense this is interesting. However, the map could in addition serve as an entry into the social science publications, starting from a specific country of group of countries. The proceeding level of clusters would then consist of the nine sub-fields. Further down the hierarchy one might find clusters of documents co-authored from groups of countries.

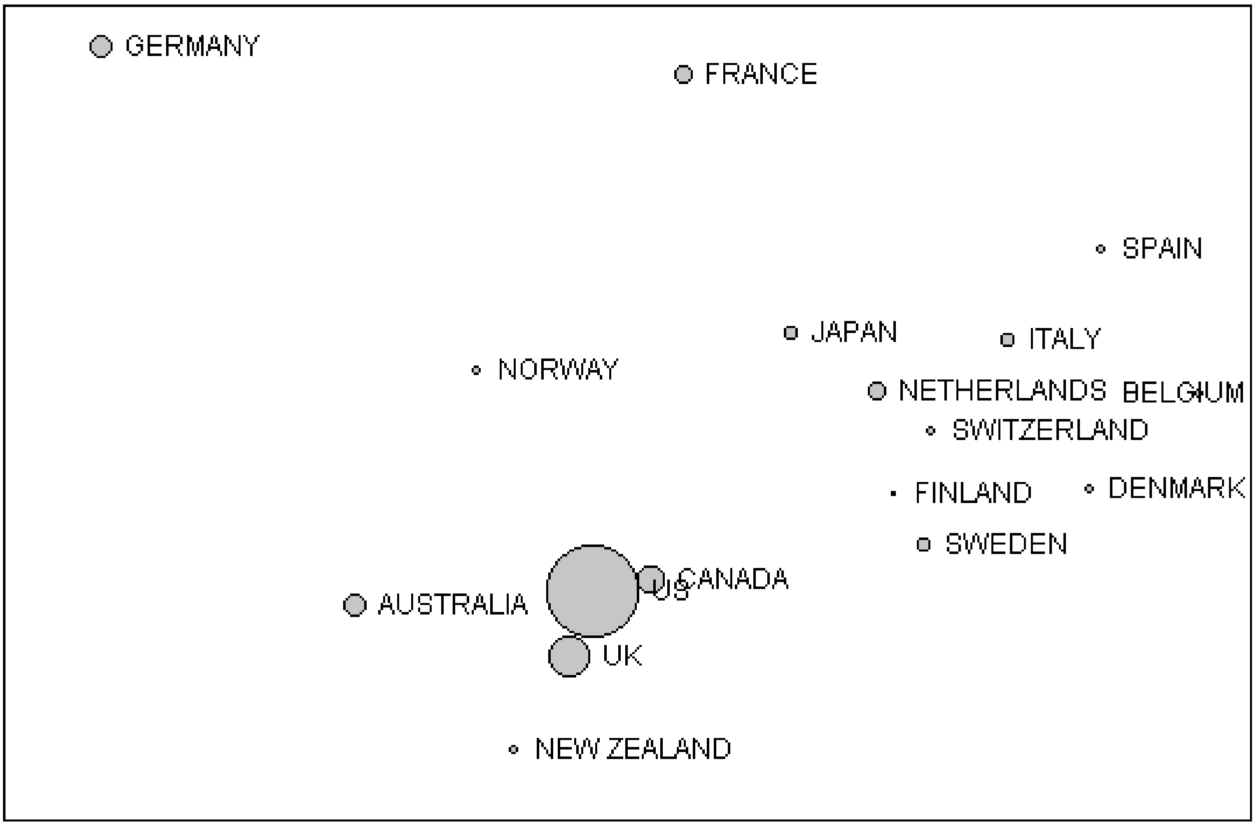

Figure 2. Publication profile map 1994-98 of 17 OECD countries covering nine social science disciplines, representing each country. Source: NSI, ISI, 1999.

3. Concluding remarks

We have demonstrated that information retrieval and informetrics are closely connected by means of the variety of representations that are available as features associated with information objects. Some of the features are of well-known nature and concerned with the aboutness of objects or documents. Other features show more descriptive characteristics. However, the analysis isolated the representations according to cognitive origin, i.e. to the type of interpretation associated with an information object. The *variety* of origin and functionality is media, domain and presentation style dependent. It is assumed useful for the purposes of retrieval and intellectual access to documents to explore this variety by means of the conception of cognitive overlaps. The more different in origin and interpretation the higher the probability that objects found in the overlap are relevant. Briefly speaking, authors and indexers are assumed to interpret the same objects slightly differently, and this difference can be intersected with the interpretations of fellow authors via received citations. The conception stresses the difference of references (or out-links on the Web) and citations (or in-links). Both mingle in a combined network – commonly known as the citation network. The conceived model of cognitive overlaps points to additional features of information objects, traditionally not associated with subject matter and topical retrieval. Such features concern journal names or affiliations, commonly seen as purely bibliographic entry points. The analysis reveals that such features are determined by more remote cognitive interpretations and assessments, for instance, by reviewing processes.

Finally the paper demonstrates that a connection exists between modern relevance types and conceptions *and* knowledge representation, seen in a cognitive and socio-cognitive view. It seems fruitful further to explore that connection.

References

Anderson, J.D. & Pérez-Carballo, J. (2001). The Nature of indexing: how humans and machines analyze messages and texts for retrieval: Part 1: Research and the nature of human indexing; Part 2: machine indexing and the allocation of human versus machine effort. *Information Processing & Management*, 37(2), 231-277.

Bates, M. (1989). The design of browsing and berry-picking techniques for the online interface. *Online Review* (now Online & CD-ROM Review), 13(5), 407-424.

Christensen, F.H. & Ingwersen, P. (1996). Online citation analysis: A methodological approach. *Scientometrics*, 37(1), 39-62.

Cosijn, E. & Ingwersen, P. (2000). Dimensions of relevance. . *Information Processing & Management*, 36, 533-550.

Garfield, E. (1998). From citation indexes to informetrics: Is the tail wagging the dog? *Libri*, 48, 67-80.

Hutchins, W. J. (1978). The subject of ´boutness´ in subject searching. *Aslib Proceedings*, 30(5), 172-181.

Hjorland, B. (1997). *Information seeking and subject representation: An activitytheoretical approach to information science.* Westport, CT, Greenwood Press.

Ingwersen, P. (1992). *Information retrieval interaction*. London, Taylor-Graham.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3-50.

Jacob, E.K. (2001). The everyday world of work: putting classification in context. *Journal of Documentation*, 57(1), 76-99.

Jacob, E.K. & Shaw, D. Sociocognitive perspectives of representation. In: Williams, M.E. (ed.). *Annual Review of Information Science and Technology*, 33, 131-185.

Pao, M. (1994). Relevance odds of retrieval overlaps from seven search fields. *Information Processing & Management*, 30(3), 305-314.

Salton, G. & McGill, J.M. (1983). *Introduction to modern information retrieval*. New York, N.Y., McGraw-Hill.

Schamber, L., Eisenberg, M. & Nilan, M. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management*, 6(6), 755-776.

van Rijsbergen, C.J. (1979). *Information retrieval*. 2. ed. London, Butterworth. (available on the net via Glasgow University, Dept. of Computing Sc. web-site)