university and society. These three projects cover technical domains as disparate as Individual and Group Behavioural, Social and Economic aspects, Networking Protocols, Interactive Browsing Technologies, Categorisation and Organisation of Electronic Information, Formal Description Techniques, Distributed Platforms, etc.

The SCREEN project (Service Creation Engineering Environment) belongs to the area of Service Engineering and deals with distributed platforms, formal methods, tools and methods for service creation. The project began in September 1996 and will define and demonstrate service creation environments (SCEs) targeted to distributed processing environments (DPEs). This project will use CORBA, the distributed-object standard developed by the Object Management Group. To evaluate SCREEN service creation engineering practices and the resulting SCE, the consortium will perform a number of trials, one of which will implement a small distributed digital library. Although, as the main objective of the project is not to build an application, certain important topics related to digital libraries will not be considered (eg property rights, interoperability, multilinguality, etc.), this implementation will bring together people and knowledge from areas as diverse as Service Engineering and Telematics Services, which are generally kept very separate.

SCREEN will deal with the whole service creation life-cycle: from the negotiation between the service subscriber and the service developer and the user (informal) specification, to formal description, modelling, implementation, testing and deployment of the service. Several languages, with different characteristics, will be used in the project, such as Visual Basic for requirements capture, Object Modelling Techniques for service analysis, formal description languages like SDL-92 for service specification, and C++, HTML and Java for implementing Web related services. The Babel (Bibliotecas Digitais como Base do Ensino Lusófono – Distance Learning based on Digital Libraries) initiative will promote the development of digital libraries in the University of Porto as a network of knowledge bases with functionalities for distance learning, self education and scientific research. This project will cover research areas such as Information Retrieval, Browsing Interfaces, Educational Methodologies and Legal, Commercial and Security aspects. The rationale behind the project is the creation of a framework for multidisciplinary R&D activities in the area of digital libraries and to promote cooperation among university faculties and research centres.

The purpose of Infocortex is the specification and prototyping of a multimedia database system with information about the available expertise in several faculties of the University of Porto. This database will provide information that can be consulted by companies or businesses. The initial idea was to create a digital library of papers and theses produced by students and staff of the University, accessible mainly by people at the university This objective was then extended to make the information available to the outside world, in a format that could be used by people not directly involved with the academic world. The project will begin with the identification of the particular information needs of companies working in different sectors, followed by the specification and systemisation of the information system of the University of Porto. Based on these studies, a prototype will be implemented that can be used to promote the system and motivate the agents involved, and which will also foster the establishment of protocols with other universities and groups of enterprises.

It is our belief that these projects will lead to advances in quite different areas but that each of them is of great importance to the topic of Digital Libraries. We expect to have the preliminary results of these projects on-line by mid 1997. ■

Please contact:
Paula Viana – INESC
E-mail: pviana@porthos.inescn.pt
or Eurico Carrapatoso – INESC
E-mail: emc@inescn.pt
Tel: + 351 2 2094220

Bipartite Graphs and Automatic Generation of Thesauri

by Michiel Hazewinkel

The key to information finding (information retrieval) in large and very large collections is METADATA, that is extra information added to the main document (article, or chapter, or...) designed to make it findable (retrievability) – particularly, metadata in the form of subject-indicators (key words and key phrases) and classification information.

A thesaurus and classification scheme can be of enormous value to make a given (large) corpus of material accessible. This is 'proved' for example by the very considerable effort that Elsevier Science Publishers invests in maintaining the thesaurus of 30,000 concepts underlying EMBASE (Excepta Medica, four full time employees). Of course there are many other aspects to the general idea of metadata. For instance, the matter of suitable containers for Metadata has attracted considerable interest in recent years.

Building a sufficiently large thesaurus by hand is an enormously expensive and laborious task; certainly if we are thinking of one of the size really needed to deal with a field like mathematics or computer science (let alone a field like physics). For mathematics I estimate that a thesaurus of key phrases (not words – these are not information rich enough) should comprise about 120,000 terms. Thus it is instinctive to begin to think about the automatic generation of thesauri. By now there are (first generation) commercial and academic programs available for extracting index and thesaurus terms (noun phrases, pre-positional noun phrases) from a corpus of electronic texts. Other linguistic software can be used to 'standardize' the raw list of key phrases thus obtained. The available

data currently consists of a bipartite graph between terms and documents. At this stage, a number of mathematical and computer science problems also arise and I should like to mention some of them.

Thesaurus problems

A thesaurus according to ISO standard 2788 and a number of other national and international standards, is much more than an alphabetical list of standardized key phrases. Fortunately, using the Hamming distance, a metric is defined on both the set of terms and the set of documents. This gives a notion of distance between terms (and between documents) thus giving a quantitative version of the thesaurus concept 'related term'. This also gives promising possibilities for such things as neighbourhood search. A much harder problem mathematically is how to capture (quantitatively) the thesaurus concepts of 'broader terms' and 'narrower terms'.

Bottom up classification

The notion of distance on the set of terms makes it possible to apply clustering techniques and thus to define a hierarchy. The problem is to relate this bottom up hierarchy with the top-down hierarchy represented by a classification scheme for a given field.

Missing terms problem

The set of terms generated as above will have a tendency to be incomplete, particularly as regards the more general, broader terms (which can also be thought of as missing centres of clusters). The problem is how to recover these missing centres.

Matching problem

Generating a thesaurus in one go for a large area is not feasible (Mathematics is as large a field as I care to contemplate in this respect; life sciences or physics are far larger fields). Thus the problem arises of constructing several thesauri (to be thought of as charts forming part of an atlas) and to match them, ie to describe the overlap between them.

These and various other related problems (such as the multilingual version of the Matching problem) form the subject matter of a number of research projects being carried out at CWI in the context of Digital Libraries. Related efforts involve interactive books and multimedia.

■

Please contact:
Michiel Hazewinkel – CWI
Tel: +31 20 5924204
E-mail: mich@cwi.nl

Information Retrieval and Metadata – Digital Library Activities at SICS

by Preben Hansen

The introduction of the Internet and WWW is bringing major changes to the role of libraries. Digital Libraries are also of growing importance in the field of Information Retrieval. At SICS, we are involved in both national and international projects in this sector and we describe two areas of activities:

Nordic Metadata Project.

This is a joint Nordic project, sponsored by NORDINFO. The project is based on the Dublin Core Element Set, which consists of thirteen metadata elements developed in collaboration by OCLC, NCSA, Library of Congress and the British Library. At the moment there are other Dublin Core-related activities in the US, Europe and Australia.

From the libraries' point of view, metadata provision poses some interesting challenges: ie growing information resources, unstable digital documents and new versions or renamed documents. The term metadata is increasingly being used in the information world to specify records that refer to digital resources available across a network. It is of major interest that embedded metadata can be utilized globally. The ultimate aim of metadata provision is to enhance end-user services by making digital documents more easily searchable and deliverable.

The Nordic Metadata Project will:

- evaluate existing metadata formats
- create a Nordic version of the Dublin Core and its DTD
- convert Dublin Core to Nordic MARC formats and vice versa
- create DC Metadata Syntax, User Environment and Interaction. This will include DC syntax requirements and recommendations; DC user guidelines; Coordination of DC test collection creation; Information Retrieval interaction and evaluation. We will adapt to the new functionality through evaluation and adaptation of the user interface, search support and support guidelines
- improve retrieval of Nordic Internet documents. The Nordic Metadata Project intends to develop and add new facilities to NWI (*Nordic Web Index* funded by NORDINFO), so that it can provide a basis for resource description, discovery and retrieval. To promote the production of metadata and the implementation of services, it is important to have a search service, capable of using the metadata. It is equally important that the search service treats metadata correctly, adapts to the standards agreed upon, demonstrates the importance of metadata for improved retrieval and supports search processes. This will be done through modifying NWI's harvesting and indexing software, by adaptations and improvements of the retrieval system and by creating test environments for retrieval experiments.

Interface Design and Information Seeking Strategies

Research in information retrieval has traditionally concentrated on building