IRSG Informer

BCS

CEPIS

Volume 3 Winter 1994

New editor, new name

And it's goodbye from him (Mark Sanderson) and hello from me (Mark Magennis), the new editor. That's my picture on the right if I've managed to get to grips with the photo retouching software by the time this goes to press. In order to indelibly stamp my authority right from the start, I've decided to change the name. No more the dreary 'BCS IRSG newsletter of the British Computer Society...blah, blah, blah'. From now on (or until the next editor changes it back again) it will have the snappy title of 'IRSG Informer'.

As the new name suggests, the primary purpose of the newsletter is to 'inform' you of what's going on in IR and in the IRSG. Our membership is very diverse, covering theoretical researchers, librarians, professional information searchers, commercial software developers, and more. The newsletter is intended to bring together the interests of this broad range of IR workers and to inform them of current developments and events across the whole scope of the field.

Mark Sanderson has done a great job over the last three years in making it an interesting mix of serious and not-so-serious news and articles. For the moment, I'm more-or-less carrying on in the same vein, although I would like to see the Informer expand its content and in future I hope to have regular columnists producing update reports in particular fields of interest. I would also like to include more product reviews and job announcements. From this issue there will be regular interviews with a prominent member of the British IR community, starting with the committee members.

Conference calls will still appear, or at least the essential details, together with whatever description I feel is appropriate or necessary and pointers to where you can get further information. The concept of 'pointers' to information sources is really taking off with the explosion of Internet use and, in particular, the World

Wide Web. I don't intend to ignore what is probably the most important development in information dissemination since the invention of the printing press (and you can quote that if you like), so you will find an increasing number of Uniform Resouce Locators (URLs) within these pages. In fact, the Informer may eventually *be* a URL.

I hope you enjoy reading this issue and won't hesitate to contact me if you have anything you'd like to include in the next.

All unattributed articles have been authored by the editor. Attributed articles may have been edited, either with the express consent of the author or in such a way that the meaning has remained unchanged.

- Mark Magennis, Editor

1994/95 IRSG Programme

You should find a copy of the BCS IRSG programme for 1994/95 included with this issue. It contains details of the 1995 colloquium and the Electronic Data Interchange (EDI) meeting, both to be held at Manchester Metropolitan University between the 3rd and 6th of April.

1993 Colloquium Proceedings published

At last, and with good looks too, the proceedings of the 1993 BCS IRSG colloquium held at Strathclyde University are being published as a book with the title "Information Retrieval: new systems and current research. Proccedings of the 15th Colloqium of the ... etc". The publisher tell me that copies are being sent to libraries, reviewers, etc. He will send one to each speaker directly, we should get batch and royalties will be sent directly to the chairman in due course. Some of the pictures did not come out too well though.

- Ruben Leon (ruben@dis.strath.ac.uk)

So who are the IRSG?

The IRSG is a specialist group of the British Computer Society (BCS) and the Council of European Professional Informatics Societies (CEPIS). It concerns Information Retrieval (IR), probably the most relevant and exciting area of scientific endeavour of the 90's.

There are currently 408 individual members and one commercial sponsor - Health Libraries & Information Network. Expect to find out more about them, what they do and why they are sponsoring the IRSG, in the next issue of the Informer.

The individual membership covers a very diverse mix of accademic and commercial work ranging over all areas of IR and peripheral or related research and development. We hope soon to compile a directory of members and their interests.

The IRSG is currently being insidiously infiltrated by a secret society of Marks from Glasgow.

Editor:

Mark Magennis
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ

email: mark3@dcs.glasgow.ac.uk
tel: 0141 339 8855 x8452
fax: 0141 330 4913

The BCS IRSG is sponsored by

Health Libraries & Information Network

The Wonderful World of the Internet

A regular column detailing the latest sources of information, or maybe just pure pap, that is out there, only a couple of mouse clicks away.

Internet Search tools

If you can only ever remember one URL then it probably ought to be http://lycos.cs.cmu.edu which gets you **Lycos**, a World Wide Web retrieval system. Lycos currently indexes the texts of over a million URLs and searches by exact or prefix matching but doesn't yet offer logical operators, except for the dreaded NOT!! Warning - there is currently no stopterms list so don't include words like 'URL' in your query.

Yahoo! is a manual indexation of the Web from http://akebono.stanford.edu/yahoo. Yahoo! hiearchically categorises URLs and currently includes over 22,500 entries. You can search for words (or substrings) in URLs, titles and/or the manually-added comments, but it doesn't index the actual content.

The Internet Research Task Force Research Group on Resource Discovery (IRTF-RD) has made its **Harvest** software system available on the Internet. Harvest is an integrated set of tools to gather, extract, organize, search, cache and replicate relevant information across the Internet. With modest effort users can tailor Harvest to digest information in many different formats, and offer custom search services on the Internet. Moreover, Harvest makes very efficient use of network traffic, remote servers, and disk space.

A number of content indexes have been built with Harvest, including an index of AT&T's 1-800 phone numbers, an index of WWW home pages, and an index of over 24,000 Computer Science technical reports from around the world.

You can get demonstrations, papers, software and documentation about the Harvest software system, from http://harvest.cs.colorado.edu/.

- Mike Schwartz, IRTF-RD Chair
(schwartz@latour.cs.colorado.edu)

Miscellaneous sources

The Telegraph newspaper is now on the Web, called 'The Electronic Telegraph' at http://www.telegraph.co.uk. It has been going since 7th November and apparently the most frequently asked question so far has been 'can we have a crossword please?'.

The "virtual" law library reference desk (VLLRD) aims to:

- provide access to GENERAL reference sources and guides
- provide access to matter that enriches academic study

As well as the Constitution of just about every country that has one, you can find such gems as the Maastricht treaty, the Magna Carta and the German WW2 surrender documents. Well I thought they were interesting anyway and there's no way I'd have got any of them by any other means. The VLLRD is at http://law.wuacc.edu/washlaw/reflaw/reflaw.html.

The one that started it all, Vannevar Bush's seminal 'hypertext' paper from 1945, 'As we may think' is now available from http://www.csi.uottawa.ca/~dduchier/misc/vbush/as-we-may-think.html. Vannevar looks like a pretty cool dude to me.

Project Gutenberg aims to get the complete text of 10,000 major books and other documents on-line by 2001. These range from the sublime 'Complete works of William Shakespeare' to the ridiculous 'The square root of 2 (to 5 million digits)'. Look in http://www.germany.eu.net/books/gutenberg/gutenberg.html.

The European equivalent of US President Al Gore's thoughts on a National Information Infrastructure is the Bangmann Commission's 'Europe and the Global Information Society - Recomendations to the European Council' available from http://www.earn.net/EC or ftp://ftp.rare.nl/upturn/docs/bangemann.ps

Even our own government is entering the information age. The CCTA Government Information Service can be found at http://www.open.gov.uk. It aims at "...achieving the aims of the Citizens Charter and enhancing open government", according to Robert Hughes MP who you can now see (looks like Oliver North to me) and hear (sounds like Sylvester the cat) in your own office. You can also access H.M. Treasury at http://www.hm-treasury.gov.uk. Expect to recieve repeated "No such file or directory" messages if you are a known left-winger or other such leather-clad subversive. It'll probably all be lies anyway. When you're fed up with one bunch of politicians why not hop over to http://www.poptel.org.uk/Labour-Party and see what the opposition have to say.

Therapeutic sources

Need romance but can't drag yourself away from your computer keyboard? Try the Internet Love Link, a UK lonely hearts service at http://www.cityscape.co.uk/lovelink.

Information and support for suferers of chronic fatigue syndrome (CFS) is available from http://www.ncf.carleton.ca/freenet/rootdir/menus/social.services/cfseir/CFSEIR.HP.html. As well as information, you can get in touch with health services and support groups.

After roaming around the Internet for a while looking for interesting and useful services I find it quite exciting that a lot of what is available is not the mindless pap you might expect. On the contrary, well-conceived services like the one for CFS sufferers aren't at all unusual. Maybe this will change as the Internet becomes more enticing from a commercial point of view, but at least for the present the content of webspace is eclectic and stimulating. The most striking thing about what can be found on the Internet is that a lot of the time you know that you wouldn't have come accross it any other way.

And finally, a cure for information overload. We all know the problem. Your spotty annorak-wearing colleague down the corridoor has just rushed in and told you about the latest 50 essential URLs that you simply must see or risk becoming one of the uninformed, disposessed multitude.

continued next page...

Internet cont...

Trouble is, you still haven't had time to look at the last 50, or the 50 before that. You're panicking and wishing to God you had a device that would freeze time for a week so that you can catch up whilst everyone else sleeps.

Face the truth, you'll never get the time to access all those 'essential' URLs. To take the brain-ache out of deciding which ones *not* to look at, use URouLette from http://kuhttp.cc.ukans.edu/cwis/organizations/kucia/uroulette/uroulette.html. URouLette generates random URLs so you don't have to think about it at all.

Some IR sources

A few of the most useful sources of information and discussion used by members of the British IR community, for those who don't already know.

ir@mailbase.ac.uk is an email disussion list used, though not very much at present, by the British IR communuity. To join, send an email containing the following message to mailbase@uk.ac.mailbase:

JOIN ir <first name> <last name>

IRLIST Digest is an electronic newsletter distributed by email from the University of California. It appears monthly and includes conference announcements, job offers, IR-related dissertation abstracts and other titbits. You can also post requests for help or information to it, reaching a great many IR workers worldwide. To subscribe, send an email request to ncgur@uccmvsa.ucop.edu. IR-L Digest is also on the World Wide Web at http://www.dcs.gla.ac.uk/scripts/global/wais/ir_list_form.

IDOM-Web is a "prototypical network document to improve the communication and technology exchange between European research and development groups with expertise in database and information retrieval". Take a look at http://idom-www.informatik.uni-hamburg.de/Idomeneus/BBoard/entry.html

Wide open govt.

Apparently, within 6 minutes of the new Government Information Service appearing on the Web someone had hacked into it and re-arranged the pages to look better.

C^{e}DAR

C^{e}DAR - the Centre for Database Access Research at the University of Huddersfield was officially formed in 1993 as a focus for research and consultancy in the School of Computing and Mathematics by Steve Pollitt (long term IRSG member and former chairman (79-84)). The Centre has focussed attention on improving user interfaces to databases, both bibliographic and data, and has been demonstrating how thesauri and classification schemes can be made central to the user's query specification.

C^{e}DAR has now completed two consultancy contracts for the European Parliament. A report with recommendations on how the Parliament can improve access to EPOQUE (European Parliament Online Query System) was submitted to the Parliament in July '94. A second report advising on the implementation of VUSE (View-based User Search Engine) techniques was delivered in October.

The consultancy for the Parliament follows directly from the construction of prototype systems on both the PC and Apple Macintosh which present a new interface and mode of searching for the EPOQUE database in Luxembourg. These systems were demonstrated to the Council of Ministers Working Party on Legal Data Processing in Brussels in May 1993, and subsequently installed in the offices of MEPs and at the Parliament's library in Brussels.

C^{e}DAR staff involved in the software development were Geoff Ellis, Martin Smith, David O'Brien, Chun Sheng Li and Steve Pollitt.

The main task for C^{e}DAR is now one of technology transfer and partnerships are being forged to further the application of VUSE techniques for databases in different subject areas. VUSE for INSPEC, developed with collaboration from INSPEC and STN International, is now the subject of an experiment at the IEE Library in London.

The principles underpinning the VUSE techniques are being presented at this years International Online Information Meeting in London harking back to the days before we started "talking" to machines - "Peek-a-Boo revived - End-user searching of bibliographic databases using filtering views".

For further details of C^{e}DAR activities contact Steve Pollitt/Ann Jones on 01484 472248 or email cedar@hud.ac.uk.

- Steve Pollitt (pollitt@hud.ac.uk)

Football more important than State reception

This issue of the Informer carries a report on the SIGIR '94 conference in Dublin. Whilst this includes much useful information about the papers and pannel sessions it fails to address the focal point of the conference - the Irish World Cup challenge. It really became clear how important this was when it was time for the official State reception with the Finance Minister. Due to a crucial Irish match, our planned soiree was postponed for two hours, with an option on a further extension should the match go into extra time. Unfortunately Ireland lost so we all traipsed off to drown our sorrows, only to find out that the Finance Minister couldn't make it since he'd decided to go to the match instead, which was in Florida. Nobody blamed him.

BCS IRSG CEPIS

How to get the most out of the World Wide Web

It occured to me, whilst jumping randomly around webspace, that URouLette could be used rather like the 'I Ching' to generate truly useful URLs. The way it works is this. You do something purposeful, (in this case you click on a link instead of heating up an old turtle shell or casting a bunch of yarrow stalks), something random turns up, and you interpret the meaning (or information) in it. The trick is to be in the right state of mind when you make the interpretation so that your innate connection with the entire universe, past and present, provides you with just the information you needed. To reach this state of mind of course requires a good deal of prior meditation and generally sitting around on your own in a cave, but that's the beauty of it. Not only do you get the information you need, but you also bannish stress, high blood pressure, mortgage payments, etc.

SIGIR '94 Dublin report

SIGIR 94, 17th International Conference On Research And Development In Information Retrieval, Dublin, Ireland 3-6 July 1994.
Report by Stephen E. Robertson and Peter Willett prepared for BLRDD
Permission to include it in the Informer was kindly given by the British Library Research and Development Department.

Introduction

The series of conferences under the title of *International Conference on Research and Development in Information Retrieval* has now been running since 1977. The annual conferences alternate between America and Europe, under the general direction of the Special Interest Group in Information Retrieval of the Association for Computing Machinery (ACM SIGIR) and with co-sponsorship from the leading European computer societies (AICA-GLIR in Italy, BCS-IRSG in the UK, DD in Denmark, GI in Germany, INRIA in France, and ICS in Ireland). The next three conferences are to be in Seattle (1995), Zurich (1996) and Princeton (1997), with initial planning for the 1998 conference focusing on Brisbane, Australia. This year's conference, the 17th, was organised by Dublin City University with financial support from Aer Lingus, Bord Failte, the Commission of the European Communities, the IDOMENEUS ESPRIT Network of Excellence and the National Software Directorate. In all, 35 papers were selected for presentation at the conference, out of a total of 129 that were submitted for consideration by the two programme committees (one for America and one for the rest of the world under Bruce Croft and Keith van Rijsbergen, respectively). The acceptance rate of less than one in three thus means that the presented papers represent the leading edge of current research in information retrieval. Almost one half of the papers came from the USA but there were also contributions from Australia, Denmark, France, Germany, Holland, Italy, Japan, Korea and the United Kingdom. There were two keynote presentations and the ACM SIGIR award paper, and two panel sessions, making this the largest such conference to date. The conference programme was divided into sessions covering Text Categorisation, Indexing, User Modelling, Theory And Logic, Natural Language Processing, Statistical Models, Performance Evaluation, Probabilistic Models, Interfaces, Routing, Passage Retrieval, and Implementation. There were also panel sessions on Integration Of Information Retrieval And Database Systems and on Evaluating Interactive Retrieval Systems. The social programme consisted of a reception at the Ministry of Finance on the Monday evening, and the conference dinner on the Tuesday evening in the beautiful Royal Hospital, Kilmainham, which was followed by a Irish folk concert that ended with the long-to-be-remembered sight of many luminaries of the information-retrieval world flinging themselves around the dance floor.

Monday 4th July

Keynote Address

The conference opened on the Monday morning with Tsichritzis's keynote address *What is Information in Information Retrieval?* He started by illustrating the changing form of information over the last few decades, as perceived and used in information retrieval, *i.e.* from numerical information to textual information to audiovisual information, and the media in which this information is stored. The speaker emphasised the extent to which the forms of information and the methods by which they are retrieved, are led or influenced by market forces. The emergence of the *World Wide Web* (especially *Mosaic*, its most popular end-user interface/browsing tool) was cited as a key example of a global multimedia information retrieval system. The major part of the presentation consisted of a comparison of the functionality of traditional databases with audiovisual databases, as well as noting some characteristics and requirements of the latter. Tsichritzis concluded by defining one of the major challenges of IR as being the integration of communication and other technologies.

Text Categorisation

Lewis and Gale (*Training Text Classifiers By Uncertainty Sampling*) described the testing, on a newswire categorisation task, of an algorithm for sequential sampling of statistical classifiers during machine learning. The author briefly explained the theoretical approach to uncertainty sampling in information systems; conventional relevance feedback relies on the user to label texts that they regard as relevant, while uncertainty sampling relies on the manual labelling of example texts, classifier fitting from these results and the re-labelling of examples whose class membership is unclear. The results of experiments with document test collections showed that, with the data used, uncertainty sampling outperformed relevance sampling. This was a very clear and well presented paper. Yang (*Expert Networks: Combining Word-Based Matching and Human Experiences in Text Categorisation*) then described the use of an "expert network", called ExpNet, automatically to categorise and retrieve natural language text. The aim of ExpNet is to deal with the vocabulary differences that exist between documents and categories, and user queries and documents. At a simple level, ExpNet consists of a three-layer network of word, text and categories. To learn, the network uses the computations of within-document word weights and the conditional probabilities of text bodies being assigned to categories, and of words to text bodies; this makes for a relatively small amount of learning and computation. The main conclusion of the testing of *ExpNet* was that it was superior to many conventional word-based matching methods in terms of both efficiency and effectiveness. Apte *et al.* (*Towards Language Independent Automated Learning of Text Categorisation Models*) then discussed machine-learning experiments on collections of English and German newswires; the aim was to discover classification patterns that could be used for the allocation of topics to specific newswires. The methodology used produced no significant difference in results when the alternative language was used, and could therefore be considered to be language independent. The overall results for the English newswires showed an increased performance over established benchmarks, while the (less complete) results for the German newswires gave promising results. The final paper of this session, by Hoch (*Using IR Techniques for Text Classification in Document Analysis*), described a working system, INFOCLAS, that took German business letters and classified them according to whether they were an order, offer, inquiry, confirmation or advertisement. The system consisted of an indexing system, which extracted and weighted the indexing terms, and a classifier. It incorporated

SIGIR '94 Dublin report cont...

several knowledge sources, including a letter database and a list of German word frequency statistics. The system produces a set of weighted hypotheses about the categories to which a letter is applicable. In the test runs, 42 letters were entered, of which 24 were assigned to the appropriate classification as the first hypothesis, and 8 were assigned to the appropriate classification as the second hypothesis. Even though these results may be unsatisfactory in a working environment, it was an interesting example of how several elements of information retrieval can be combined to produce a functioning system. A great attraction of the system was its processing speed, taking between 0.5 and 2 CPU seconds to classify an entire letter.

Indexing

Two of the three papers in this session addressed questions relating to the measurement of the effectiveness of parts of a retrieval system, rather than the performance of the whole (as in the traditional retrieval test). This is a general problem in information retrieval, in part because one would like measurements to reflect very directly the aspect of the system which is under consideration; whole-system performance measurement tends to obscure the effects of particular parts and to make diagnosis difficult. Indexing, in particular, suffers from this problem. However, any attempt to measure the effectiveness of part of the system is likely to face the problem of determining whether and how a "good" result actually relates to good retrieval performance. Paice (*An Evaluation Method for Stemming Algorithms*) made this case, and presented a criterion method whereby stemming algorithms are assessed against an ideal set of conflated groups of terms. The method provides measurements of understemming and overstemming, and may be used diagnostically in respect of particular rules in each stemmer. Of the three stemmers used in the experiments, Lovins (the oldest) seemed to be the least accurate overall, Porter is a light stemmer (tends to understem), while Paice/Husk is heavy (tends to overstem); however, Paice did not try to demonstrate any relation between his measurements and whole-system performance. Furner-Hines *et al.* (*On the Measurement of Inter-Linker Consistency in Hypertext Databases*) drew a parallel with inter-indexer consistency studies and described an experiment which measured inter-linker consistency and also retrieval effectiveness. Consistency on the whole was low; on the other hand, they were unable to show any link between consistency and retrieval performance. We are left with the question (which indeed remains almost equally open for indexing) as to what characteristics of the linking process affect retrieval performance and in what way. In the third paper, which might have fitted better in a session on query expansion rather than indexing, Voorhees (*Query Expansion Using Lexical-Semantic Relations*) discussed the use of WordNet synonym sets for this purpose. WordNet is a general-purpose, lexical-semantic structure bearing more similarity to a thesaurus of the Roget type than to one designed for IR. An experiment on the TREC collection revealed little overall benefit, albeit with some evidence of benefit on shorter queries (the TREC queries are highly detailed, carrying far more information than is typical for, say, online searching).

Panel session: Integration of Information Retrieval and Database Systems.

This panel session was run in parallel with the session on Indexing, and was opened by Fuhr, who summarised the goals of integrating the exact match of database systems and the partial match of text retrieval systems as providing, in one system: a uniform query language; probabilistic document indexing; search term weighting; searching of imprecise data; and searching of vague facts. He described a probabilistic NF2 model which includes nested relations and probabilistic tuple weights; weights can be given to terms or authors, (higher to the first author), and these can be combined in retrieval, and also incorporate Boolean retrieval in arbitrarily complex queries. There is an interface between NF2 and SQL, and the output is a ranked list. Larsen described Postgres, an Object-Oriented Database System. The problem arises from the storage of very large databases as single attributes, which leads to very large storage requirements, with separate indexes and external programs for access. Indexing sub-elements of objects requires complex operations. This work is in a very early stage as far as implementation is concerned. Schäuble described Spider, a probabilistic IR system that is available *via* the World Wide Web. It incorporates weighting and relevance feedback as well as a complete set of DBMS features. While this offers potentially simple integration of DBMS and IR functionalities, the problems include: ensuring data consistency if there are frequent updates of the IR data; combining IR and database queries; and fact retrieval. Schmidt then described the difficulties of using SQL. Although it was originally intended as a unifying architecture combining a data model, a query language and a database system, 25 years of development have resulted in 600 pages of extensions to all three parts. It is now thus not a standard; instead, "SQL is a moving target". Finally, Thiel looked to the future when DBMS and IR systems may converge into one, rather than existing as hybrid systems. Three options are possible: integration; hybrid, or loose-coupled, with both available; combine features into a new system design. He then mentioned Hydra which does the first by means of Sybase and Inquery, Cordis, which uses a fill-in-the-blanks screen, and the MIND project for multimedia integration.

User Modelling

Four papers were presented in the session on user modelling: two based on empirical studies and two addressing more theoretical issues. Allen (*Perceptual Speed, Learning and Information Retrieval Performance*) looked at how perceptual speed, *e.g.*, speed in scanning subject descriptors, could influence vocabulary learning and retrieval performance. One hundred subjects were given the task of retrieving related relevant documents from a small test collection of 256 references after reading an article on aggression and completing two independently scored tests of perceptual speed. Two versions of the system on CD-ROM were randomly assigned: one presented subject descriptors for each reference as a first element in the display and the other presented them in the normal way after author, title and source. Dependent variables included: precision and recall, amount of vocabulary learning and time in completing the search. The main results showed a significant interaction between the order of presenting subject descriptors and users' perceptual speed in retrieval performance measured by recall and precision, in the amount of learning of search

SIGIR '94 Dublin report cont...

vocabulary as well as in the speed in carrying out a search. That is to say subjects who could quickly scan subject descriptors could extract more appropriate descriptors to achieve better search performance and work more quickly. The findings indicate the importance of cognitive aspects and interface features in the retrieval task and the need to take account of individual differences and usability issues in system design.The study by Spink (*Term Relevance Feedback and Query Expansion: Relation to Design*) was also concerned with how systems could support users in search formulation. Forty mediated searches were used to examine the selection and effectiveness of search terms sources for query expansion. The objective was to ascertain the effectiveness of terms obtained interactively through retrieved documents (Term Relevance Feedback) compared with other sources including: terms from users' written search statements, user suggestions prior to the online search, terms suggested by the intermediary prior to the online search, and thesaural terms. Although TRF accounted for only 11% of search terms and contributed to only 9% of the retrieval of relevant items per search, a high proportion of TRF were productive but not as productive as user generated terms. Intermediaries identified the greater proportion of TRF and selected more effective terms, the majority of which were extracted from the title and descriptor fields and not from abstracts. The author went on to suggest that automatic relevance feedback could be implemented to provide a higher weight for terms selected from title or descriptor fields and to allow for greater user participation in term selection. Logan *et al.* (*Modelling Information Retrieval Agents with Belief Revision*) described the implementation of an intermediary simulation based on the integration of two theoretical models: the belief revision model of Galliers and the functional distributed expert model of Belkin *et al.* The work is an initial attempt to demonstrate the accuracy of the models and the long-term viability of the approach. Both models had to be extended to incorporate intentions, planning and inference and to allow for overall dialogue management. The main feature was that the blackboard architecture with its multi-agent approach for the intermediary was abandoned in favour of a single control module and distinct specialised rule sets. The focus was thus on the development of dialogue model based on a schemata of primitive speech acts. The modelling exercise identified a number of problems including: computational complexity, connectivity, predictability, communicating commitment and focusing in dialogue. In its attempt to apply theoretical concepts the study marks an important step for the development of an interactive dialogue for an intelligent intermediary and was one of the more innovative and exciting papers of the conference. The final paper in the session (*Polyrepresentation of Information Needs and Semantic Entities: Elements of a Cognitive Theory for Information Retrieval Interaction*) was presented by Ingwersen. He put the case for extending the representation of information needs and semantic entities in a global cognitive communication model of IR. The basic argument is that the uncertainty and unpredictability characteristics inherent in IR interaction require the provision of intentional redundancy both in the system's information space and the user's cognitive space. The interface between the system and the user becomes a request model builder which supports interest descriptions, the development of problem goal statements and different request versions. The model thus adopts a cognitive evolutionary approach to information need as opposed to the traditional reductionist view leading to a query. It was suggested that the model could be implemented in Okapi, building on the system's existing interactive relevance feedback and query expansion features.

Tuesday 5th July

Theory And Logic

This session contained two submitted papers and one of the three invited papers. The two submitted papers both start from van Rijsbergen's logic-based approach to IR. Bruza and Huibers (*Investigating Aboutness Axioms Using Information Fields*) proposed a logical framework that claimed to allow IR systems to be compared in terms of the implicit assumptions governing them (as opposed to an empirical comparison). In this method, each IR system or mechanism is described in common logical terms, and the associated logical axioms are identified. The framework is, however, very narrow and restrictive in several ways. For example, it assumes a general category of "information carriers" that includes both index terms and documents (with no distinction or discussion of the senses in which either might be said to "carry" information); it assumes a notion of "aboutness" as a dichotomous logical variable, again subsuming a number of different notions; it deals only with logical structure, making no concessions whatever to any cognitive aspects of IR. This last point is perhaps inherent in the van Rijsbergen approach.Sebastiani (*A Probabilistic Terminological Logic for Modelling Information Retrieval*) developed a specific interpretation of the van Rijsbergen approach. It may be said to suffer from the last problem above, though in this case the problem is in some degree mitigated by the incorporation of probabilistic notions (which is the main point of this paper). Sebastiani distinguished between probabilistic information deriving from degrees of belief and that resulting from statistical information; these play rather different roles in the model.In the invited talk, Carbonell (*Beyond Keywords; the Case for Natural Language Processing in Extended Information Retrieval Systems*, which is not published in the proceedings) gave an enjoyable and revealing short history of NLP. The core of his argument was that NLP wasted a lot of effort on highly elaborate forms of analysis that were supposed to be general purpose, but has recently (specifically in machine translation) taken a much more task-oriented approach; this has proved to be effective in MT and may also be so in IR. This presentation led neatly into one of the next sessions, which was devoted to NLP and which was run in parallel with the session on statistical models.

Natural Language Processing

The first paper of this session was by Jacquemin and Royaute (*Retrieving Terms and their Variants in a Lexicalized Unification-Based Framework).* They discussed the need for automatic term-extraction procedures that can take account of the variations in word forms that are to be expected in free-text databases. They have developed a tool, FASTR (for FASt Term Recogniser) to handle the identification of basic terms and also a parser to encompass their variations. Terms are tagged with a single syntactic category using an online dictionary without attempting to resolve ambiguities, and their possible variations are defined by means of a series of logical rules. The parser

SIGIR '94 Dublin report cont...

mixes lexical identification with local syntactic analysis. Morphological analysis is used for segmentation and stemming; there is then a syntactic parsing that invokes the rules that have been activated during the stemming phase, and negative metarules are finally used to reduce the number of spurious analyses that are produced. Sanderson (*Word Sense Disambiguation and Information Retrieval*) discussed the belief that word sense ambiguity is a cause of poor retrieval performance and that performance will increase if words can be correctly disambiguated. Some elegant experiments were described in which additional sense ambiguity was introduced into a document test collection, to allow some degree of control over the level of ambiguity that was present. The results of probabilistic searches confirmed the "common sense" view of the ability of collocation to resolve word-sense ambiguity, which implies that word sense disambiguation is likely to be of benefit only for short queries. The experiments also suggested that performance is less sensitive to ambiguity than it is to erroneous disambiguation.The final paper of this session (*A Full-Text Retrieval System with a Dynamic Abstract Generation Function*) was by Miike *et al.*, and concerned the development of the system BREVIDOC (for Broadcatching System with an Essence Viewer for Retrieved Documents). This Japanese-language, full-text retrieval system enables the user to specify an area within the text from which an abstract is to be produced and also to control the size of the abstract. The system is domain independent, document structure being analysed using linguistic clues such as connectives and anaphoric and idiomatic expressions. Document structure analysis is performed in advance and is stored so that abstracts may be produced in real time. Four major processes are performed by the document structure analyser: document organisation analysis to find the headings and scope of each section of the document; sentence analysis using a morphological analyser; text structure analysis of the body of each section; and final semantic-role extraction. The selection of sentences for abstraction is based on the relative importance of rhetorical relations among the sentences.

Statistical Models

This session focused on ranking issues. Aalbersberg (*A Document Retrieval Model Based on Term Frequency Ranks*) considered the possibility of replacing the usual form of indexing in the vector-space and other weighting systems (in which an explicit weight is stored for each document-term pair) with storage of a ranked list of terms for each document with no weights. By exploiting the Zipf law, one can devise a document-query matching formula which uses the rank of the term, without the actual frequency that gave rise to it. The method was shown to give comparable performance to the vector space model. In some circumstances, the method can be used to save indexing space. Bartell *et al.* (*Automatic Combination of Multiple Ranked Retrieval Systems*) are one of a number of groups working on combining evidence from different sources in arriving at a ranked output of documents. They proposed a method for automatically optimising this process on the basis of a training set of queries; there is some similarity here to the regression methods being used by other groups to optimise performance given a number of clues. They also discussed the problem of measuring/optimising the performance of systems on the basis of user preference statements between documents, as opposed to document-by-document relevance judgements. Lee's paper (*Properties of Extended Boolean Models in Information Retrieval*) might perhaps have fitted better in the theory and logic section. He analysed the logical properties of those extensions of Boolean logic which allow ranked output, for which several such schemes have been proposed in the last few years. He identified some problems that characterise most of these methods: in particular, there is usually a problem with the normal Boolean rule of associativity. Overall, the *p*-norm method of Salton *et al.* was the one that dealt best with the various problems.

Performance Evaluation

The session on performance evaluation included three papers. Hersh *et al.* (*OHSUMED: an Interactive Retrieval Evaluation and New Large Test Collection for Research*) described a series of experiments which compared end-user searching of a sub-set of MEDLINE with searches by experienced intermediaries. In the first part of the study 101 original searches were carried out by physicians on the vector-space Knowledge Finder system (KF). These searches were then replicated by two other physicians and two librarians on the Boolean ELHILL system using both MESH and text words, and text words only. The second part of the study compared the five searches with different runs on the SMART system which included different weighting approaches and relevance feedback. The results showed that end-users achieved significantly higher recall searching KF and that recall was also significantly higher for searches by librarians using MESH and text words than by physicians using the full features. Precision was in turn significantly higher for each of the non-KF groups. Over half of the retrieved references were retrieved by only one of the five searchers. The comparison between the interactive searchers and SMART runs was not statistically significant. It is not surprising that end-users could search effectively on a non-Boolean system and that intermediaries seemed to be able to benefit more from the use of MESH than the physicians. More importantly the study highlights the difficulty of attempting to compare interactive and automatic searching and the limitations of the recall and precision measures. Although the creation of a new test collection was welcome, the study did not explore the nature of interactive searching and implications for evaluation methodology. The paper by Taghva *et al.* (*Results of Applying Probabilistic IR to OCR Text*) was concerned with testing retrieval performance for OCR-generated data. A test collection of 674 OCR documents was generated by three different OCR devices. Retrieval tests using the INQUERY retrieval engine were carried out to ascertain the quality and reliability of the OCR text, which had undergone some manual correction as well as automatic post-processing. The worst collection had the highest average precision whilst the best collection had the lowest. Precision was affected by the ability of the device to differentiate between text and graphic data as well as term accuracy which in turn had an effect on the term frequencies in the documents. Post-processing always improved precision. Probabilistic IR is thus less resistant to OCR irregularities than a Boolean system in which errors could be overcome by redundancy in the document text. To date OCR devices have been tested at the character or word level only,

SIGIR '94 Dublin report cont...

but the retrieval tests make it possible to evaluate at the document or collection level. The third paper was by Turtle (*Natural Language vs. Boolean Query Evaluation: a Comparison of Retrieval Performance*). Two full-text collections, The Federal Cases and a random sample from West Federal Supplement database, were used to run 44 Boolean queries which were generated from natural language statements. A number of retrieval engines were used to search the natural language statements including INQUERY, SMART and Personal Librarian. The top twenty documents retrieved were rated for relevance and the judgements were pooled as in TREC. Four measures were used: precision at standard recall points, raw precision and recall at a fixed rank cutoff, the relative performance of individual queries in the test set and precision at all ranks up to some maximum value. In terms of recall and precision, natural language queries performed significantly better than Boolean queries in both collections. On all the evaluation metrics used, natural language searches consistently produced better rankings. The natural language queries were particularly effective for the large collection with three out of every four documents in the top twenty ranks being judged as relevant. The main interest in the first and third papers lies in the evaluation approach and in particular the way queries are generated, and the choice of independent variables. An attempt is being made to combine laboratory type experiments with an operational setting. Even though the focus remains on intermediary or expert searchers within specific subject domains, the ultimate aim is to inform on end-user searching.

Probabilistic Models

Gey (*Inferring Probability of Relevance Using the Method of Logistic Regression*) presented some of the work undertaken in his doctoral dissertation, in which the method of logistic regression was applied to information retrieval in the calculation of probabilities of relevance of documents with respect to queries. A brief description of logistic regression was given, followed by results obtained when this was applied to the Cranfield data. Results were compared using the logistic regression model with the vector-space model, which the author proposed as a benchmark for such comparisons. Standard measures of performance were compared using non-parametric tests. The results of this analysis were then applied to two other test collections (CACM and CISI), although statistical tests did not show the same improvement. Therefore, problems of over-fitting the model to one particular data collection were addressed, and a method of transferring the results from one collection to another proposed. This had the basic assumption that the variables used in the regression model would have the same underlying distributions in all data sets, but with varying means and variances. This did lead to slightly improved performance for the CACM collection, although slightly worse performance for the CISI collection. One reason suggested for this was the unusually high number of relevant documents in this collection. Finally, the author looked at the calibration of the probability estimates, by dividing the collection of deciles of equal numbers of query-document pairs, and comparing the number of expected relevant documents with the actual number of documents. Based on this analysis, the logistic model did not appear to give a "good fit". Robertson and Walker (*Some Simple Effective Approximations to the 2-Poisson model for Probabilistic Weighted Retrieval*) proposed a method of weighting search terms within the context of a probabilistic model, whereby the evidence given by within-document frequencies and document length can be combined with the Robertson-Sparck Jones weights. Three types of model are differentiated: (a) formal models which have an underlying theoretical base, but suffer from problems of parameter estimation; (b) *ad hoc* models which allow researchers to try different variables; and (c) regression-type models which may be seen as a combination of types (a) and (b). This paper proposed using the formal models to suggest simple weighting formulae. It was stressed that the complexity of probabilistic formulas arises not from complex-looking formulae, but from having variables that cannot be calculated directly. In this approach the 2-Poisson model was taken as the starting point. Formulae were suggested that approximated the 2-Poisson weights. Results of experiments were reported using the modified formulae that gave consistently better performance than formulae taking into account only the presence/absence of search terms. These experiments were undertaken on the TREC collection. This paper indicated a way of increasing retrieval performance, using complex weighting formulae to suggest much simpler formulae, and showed that this could be effective, but did not attempt to state definitively what these formulae should be. Cooper (*The Formalism of Probability Theory in IR: a Foundation or an Encumbrance?*) defined a scale for looking at the theoretical basis of IR models, and stated that there is no such thing as a full theory in IR currently. On this scale, models such as the vector-space model, described as "geometrically inspired *adhocery*", are placed at the bottom, and the probabilistic models are placed a little nearer the theoretical end of the scale. The advantages of finding a theoretical approach were considered as were the reasons why models with greater inferential powers do not in general perform significantly better than *ad hoc* models, and whether in fact, it is worth the effort of pursuing a theory, given such results. The author seemed to have decided in favour of a theoretical approach, whilst signalling some of the pitfalls that have befallen such models in the past, *e.g.*, inconsistent theories and misstated assumptions. He also acknowledged the role that the *ad hoc* models have contributed, in allowing researchers to follow up ideas unencumbered by any theoretical considerations. In the future, it is expected that theories with more inferential power will emerge. Meanwhile, it remains unclear whether the formalism of probability theory in IR is a benefit or a burden.

Wednesday 6th July

Interfaces

Hemmje *et al.* (*Lyberworld - A Visualisation User Interface Supporting Fulltext Retrieval*) discussed Lyberworld, which is a three-dimensional navigational interface to fulltext IR systems. Lyberworld uses an hierarchical tree structure, and acts as a front-end to INQUERY (which is used as the underlying retrieval engine) and the VIBE system (which is used for assessing relevance). The user interacts with *navigation cones*, which unfold alternating tree levels of term and document views. By selecting an item, the user can display all of the paths from the selected node in a new subtree. A *relevance sphere* provides a way of seeing the relevance of particular documents: the closer a document is to the border of the sphere, the

SIGIR '94 Dublin report cont...

more relevant it is in the whole query context. Users can follow the loops in the system as often as they wish, and can explore the same paths again if they require. Whatever the benefits from this kind of system, it cannot perform better than the effectiveness of the underlying retrieval engine; however, the system does seem to give the user more control over the direction of the search, as was demonstrated by an impressive videotape that accompanied the talk. Conrad and Utt (*A System for Discovering Relationships by Feature Extraction from Text Databases*) discussed the Association System, which focuses on the recognition of domain-specific features in a textual database and on the identification of relationships between those features. The two features extracted in this application were company names and personal names, and the experiments used the Wall Street Journal database. The results of their tests suggested that feature extraction can be quite accurate and that the relationships generated are reliable. Besides recall and precision, an additional evaluation measure was used that focuses on the usefullness of the relations generated. The authors concluded by noting that their techniques could be applied to other domains such as medical or legal data systems.

Routing

Morita and Shinoda (*Information Filtering Based on User Behaviour Analysis and Best Match Text Retrieval*) from the School of Information Science in the Japan Advanced Institute of Science and Technology discussed selective dissemination of information (SDI) using filtering based on feedback derived from the user's reactions to the system's output. Some of the findings appear rather obvious, for example the fact that users spent more time reading articles rated as "interesting" than they did others, but one part of their paper was of considerable interest, this relating to the use of a substring indexing method based on digrams, (all searching is done in Japanese). Words in articles deemed to be interesting are split into substrings, and these are then used as search keys for the retrieval of subsequent articles, these searches achieving a high level of precision. Hull (*Improving Text Retrieval for the Routing Problem Using Latent Semantic Indexing*) compared latent semantic indexing (LSI) with the vector space model (VSM) for the routing problem, and it was found that there was only a slight difference for the classification of news items. LSI is far more demanding of computational resources than VSM but is better able to preserve term associations. The combination of LSI with linear discriminant analysis, to yield a classification approach called Text-based Discriminant Analysis (TDA), yielded results that were superior to both LSI and VSM, given a sufficient number of known relevant documents. The final paper in the session was by Buckley *et al*. (*The Effect of Adding Relevance Information in a Relevance Feedback Environment*), who discussed the effects of varying two sources of added relevance information in the TREC experimental setup. The first was the number of relevant documents retrieved by an initial SMART search and the second was the number of terms occuring in the relevant documents used to expand the initial query. The expanded query was weighted by using a modified Rocchio relevance feedback approach. The results suggested a direct relationship between the log of the number of terms added to the initial query and its effectiveness. A similar relationship was also observed with the log of the number of relevant documents used. The overall effectiveness increase due to relevance feedback was in the range 19-38%, depending on the number of known relevant documents. Finally, the two runs using relevance information from many systems showed that multi-system relevant documents perform slightly better than an equal number of single-system relevant documents.

Passage Retrieval

The session on "Passage retrieval" addressed an area of considerable current interest, given the existence of larger and more heterogeneous collections of full-text material. The first paper was by Callan (*Passage-level Evidence in Document Retrieval*), who used INQUERY to implement a system which retrieved whole documents, but making use of evidence at the passage level. Essentially the idea was to combine a match value for the best-matching passage with one for the whole document. Experiments with various test collections, ways of identifying passages, and ways of combining evidence were described. Passage-level information appears to improve retrieval performance substantially. Wilkinson (*Effective Retrieval of Structured Documents*) considered the situation where the documents contain explicit internal structure (as indicated by, for example, SGML), and described experiments to retrieve sections rather than whole documents, as well as experiments more comparable to those described by Callan. Again, performance benefits were obtained. Mittendorf and Schäuble (*Document and Passage Retrieval Based on Hidden Markov Models*) described a model which is capable of identifying passage boundaries automatically. The model considers documents to be generated by a stochastic process, with transitions between states (passage boundaries) identifiable by a best-fit procedure. Again, some experimental support was provided, albeit on a rather small scale.

Panel Session: Evaluating Interactive Retrieval Systems

This panel session was run in parallel with the session on Passage Retrieval, and commenced with an introduction from Dumais, who identified three principal issues as worthy of discussion: rationalisation (the importance of studying the interaction between information retrieval systems and their human users); evaluation (the measurement of the performance of retrieval systems whose mechanisms involve a high degree of user–system interaction); and generalisation (the application of findings from evaluative experimentation to the design of effective systems). She continued by contrasting the boundary that is traditionally drawn between the IR system and the outside world with that which may define modern interactive systems; the latter have interfaces that encompass the user's knowledge, goals and environment, as well as the interaction between the user and the retrieval engine. Belkin expanded on these differences by discussing the consequences that such differences have for evaluative tasks. Traditional approaches to evaluation have identified the retrieval engine as the primary actor in the system, and the goal of the system is recognised to be the retrieval of all and only those documents in the system's database that are relevant to each query presented by the user. The assumption is made that there exist single, definable, static queries, the individual responses to which may be evaluated in terms of the standard measures of recall and precision. Such a model bears little resemblance to the 'real world'

SIGIR '94 Dublin report cont...

of present-day, highly interactive IR systems, in which the user is the primary actor and the goal of the user is the retrieval of a set of documents whose number and quality are sufficient to meet their individual needs. Belkin concluded by making two principal recommendations for the evaluation of interactive IR systems: firstly, that evaluation should be carried out of samples of tasks in which real users are in real problematic situations; and secondly, that the size of such samples should be maximised. Dumais stated that interface design involves consideration not just of presentation (such as choice of fonts, layout, *etc*.), but also: support for query construction, and for different retrieval strategies; alternative methods of viewing, exploring and using retrieved information; and the integration of search and navigation. Like Belkin, Dumais noted the importance of retrieval engines in traditional approaches to IR evaluation; here, substantial changes in the engine often result in only minor changes in system effectiveness, whereas even seemingly superficial changes to the interface often lead to differences in performance of 25% or more. As an example, Dumais described her involvement in the development of SuperBook, a system allowing the user to view and browse through the full-text content of hierarchically structured documents. The facilities for browsing provided in SuperBook (such as a dynamic 'fish-eye' browser) were developed on the basis of extensive psychological testing of users' interactions with hypertext. In the course of a development cycle that progressed through three distinct versions of the user interface, the average time taken by users to complete the same predefined task was reduced from 7 to 4 minutes. Hancock-Beaulieu described the philosophy underlying the design of OKAPI, an interactive IR system aimed at untrained users that uses a library OPAC as its test database. It was intended that the interface should (i) be self-explanatory (specifically providing no online 'Help' facility), (ii) allow natural-language input and interactive dialogue, and (iii) make all functionality automatic and invisible to the user. Moreover, it was the developers' belief that the modularity of OKAPI's design would make it especially useful for evaluative purposes, where repeated minor modifications might need to be carried out. In addition, the establishment of OKAPI as a fully-operational system in a library, rather than as an experimental system in a laboratory, meant that the behaviour of real users with real queries could be studied over a long period. As an example, she discussed a comparison that had been undertaken of the use of automatic query expansion (AQE) and of interactive query expansion (IQE). The AQE facility was used in 31% of searches, and 50% of those expanded searches led to new references being retrieved; corresponding figures for the IQE facility were 11% and 31% respectively. As a result, it was suggested that the interface might be redesigned to encourage the use of AQE, and to allow the user to make active selections of query terms. Finally, Borgman described the methodology and results of the five-year Science Library Catalog Project (SLCP), the aim of which was to study the information-seeking behaviour of 8–12-year-old children. The test database used in the SLCP study was a set of simple statements about scientific terms. A task was set for users that involved identifying, from a mixed set of terms, categories of those that the user decides are related. To enable users to acquire the knowledge needed for successful completion of this task, an interface was developed and improved iteratively on the basis of evaluation in many different contexts. The interface allowed the user to adopt two different sorts of retrieval strategy, one based on recognition and browsing through hierarchies of options, the other based on the selection of keywords and Boolean matching. Analysis was undertaken of system performance (as measured by the children's success in the categorisation task), searching and navigation patterns, and exploration time. The principal research questions were whether these variables were related to the children's differential use of the two search methods, and whether they were related to individual differences such as age, gender, domain knowledge or computing expertise.

Implementation.

This final session focused upon the efficiency of IR systems, rather than their effectiveness as in most of the other presentations. The first paper was by Shoens *et al.* (*Synthetic Workload performance Analysis of Incremental Updates)* who have evaluated a range of algorithms for updating inverted files, which underlie the overwhelming majority of current text-retrieval systems, both Boolean and non-Boolean, owing to their rapid speed of response. However, this is achieved only at the expense of very substantial updating costs, since inverted-file postings lists have to be modified for each and every index term in a new document that is being added to the database (or being modified or deleted in the case of non-archival systems). The algorithms discussed here bring about reductions in the cost of updating by means of a differential strategy, based upon the length of (*i.e.*, the number of document identifiers in) a postings list in which all lists are initially considered as 'short' and held in main memory; then, as individual lists start to grow, the longer lists are progressively moved to backing storage. The efficiency of the algorithms was evaluated using both synthetic data and 64 days of *NetNews*. The design of novel best-match searching algorithms based on inverted-files has been the subject of study for over a decade. Early algorithms were deterministic, in that they guaranteed the retrieval of the nearest neighbours of a natural-language query; however, there has also been interest in the use of probabilistic algorithms in which a faster speed of response is obtained at the possible cost of missing some of the true nearest-neighbour documents. Persin (*Document Filtering for Fast Ranking*) described a probabilistic upperbound algorithm for this purpose. His procedure operates by predicting at a very early stage in the processing those documents that have a high probability of achieving a high degree of similarity with the query, and then building the in-core data structures used by the conventional inverted-file best-match algorithm only for the records that have been so identified. The algorithm appears to be extremely efficient, achieving reductions of up to 80% in response times and of up to 98% in storage costs when compared with the conventional inverted-file algorithm. It does this without loss of effectiveness; indeed, rather counter-intuitively, its use sometimes results in an increase in retrieval effectiveness, rather than the decrease that has characterised previous probabilistic algorithms.The final paper of the conference was by Anick (*Adapting a Full-Text Information Retrieval System to the Computer Troubleshooting Domain*), who discussed work carried out at DEC on the provision of computer support for a help-desk, where users ring in with software and/or hardware problems. The queries typically contain only one or two words, and these are often acronyms, system component numbers, operating system

SIGIR '94 Dublin report cont...

commands and the like (*e.g.*, "110Mbyte SCSI drive" or "sys$check_access"), that are far less detailed than the natural-language queries that characterise most IR environments. This interesting paper discussed the need for domain-specific techniques for the expansion of these initial, highly terse queries. Initial work focused on the use of computational-linguistics techniques, but the sheer size of the system (DEC's Customer Support Centers have access to some 200,000 technical documents) motivated the use of extended analysis of existing system logs to identify synonyms and word variants that had been employed for previous queries. Routines were also developed to handle systematic forms of word variation that are characteristic of the particular sublanguage that is being handled, *e.g.*, sequences such as "SYS$UNWIND", "$UNWIND_S", "$UNWIND_G", "UNWIND_S", *etc*.) and to parse compound tokens such as "vms 5.0" and "v5.0-a". These various routines have all been encoded so that they can be used for the expansion of new queries.

- Stephen E. Robertson (ser@is.city.ac.uk)
- Peter Willett (p.willett@sheffield.ac.uk)

...and next: SIGIR '95

SIGIR '95, 18th International Conference on Research and Development in Information Retrieval.
The Sheraton, Seattle, Washington, USA
July 9 - July 13, 1995

SIGIR '95 is an international research conference on information retrieval theory, systems, and applications. The ACM SIGIR conference occurs annually, alternating between locations in North America and elsewhere (e.g., Europe). This conference will interest a broad spectrum of professionals including theoreticians, developers, publishers, researchers, educators, and designers of systems, interfaces, information bases, and related applications.

The following list of topics covers the areas of particular interest.

IR fundamentals

Types: text, hypertext, multimedia (including audio, images, video)
Representations: source, conversions, storage, presentation
Information structures, interaction, time-based issues
Processing: indexing, analysis, compression, retrieval, rendering, publishing
Systems: design, implementation, measures, evaluation, architectures, scalability, integration with DBMS
Theories and models, evaluation
Reasoning: logic, case-based
Standards: SGML (and HTML), HyTime, MPEG, Z39.50, HTTP

Users and IR interaction

Modeling, empirical studies
Interface design, human-computer interaction, visualization
IR tasks, including query formulation and expansion
IR and information seeking behavior

IR and cognitive approaches

Natural language processing, linguistic resources, multilingual systems
Knowledge bases and their use
Learning: genetic algorithms, neural nets
Pattern matching, uncertainty, data fusion

Dedicated IR applications

Digital libraries: architectures, prototypes, studies, issues
Networked information (e.g. WAIS, WWW): infrastructure, tools, systems, protocols, collections, interfaces, case studies, intellectual property rights

Education in IR

Curriculum, training, tools, systems

Paper submissions by January 6th 1995. Detailed information regarding submissions is available via anonymous ftp from ftp.u.washington.edu (/public/sigir95/cfp). Questions should be addressed either to the Conference Chair or to sigir95@u.washington.edu.

- Raya Fidel, Conference Chair (fidelr@u.washington.edu)

TREC-4 Call for participation

TREC-4
January 1995 - November 1995
Conducted by National Institute of Standards and Technology (NIST)
Sponsored by Advanced Research Projects Agency / Software and Intelligent Systems Technology Office (ARPA/SISTO)
The deadline for participation applications is Jan 16th 1995
The conference will take place at NIST in Gaithersburg, Maryland, USA during Nov 1-3 1995.

The Text REtrieval Conference (TREC) has had a very successful three years and we would like to invite you to submit a proposal for participation in year four (TREC-4). The goal of this conference is to encourage research in text retrieval from large document collections by providing a large test collection, uniform scoring procedures and a forum for organizations interested in comparing their results. Both adhoc queries against archival data collections and routing (filtering or dissemination) queries against incoming data streams are being tested. The conference has grown from 24 participating systems in 1992 to 33 participating systems in 1994, with proceedings published each year, and is now the major experimental effort in the field. Dissemination of TREC work and results other than in the (publically available) conference proceedings is welcomed, but the conditions of participation preclude specific advertising claims based on TREC results.

Participants will be expected to work with approximately a million documents (2 gigabytes of data), retrieving lists of ranked documents that could be considered relevant to each of 100 topics (50 routing and 50 adhoc topics). NIST will distribute the data and will collect and analyze the results. As before, the workshop will be open only to participating systems that submit results and to government sponsors.

Participants will receive 3 gigabytes of data for use in training of their systems, including development of appropriate algorithms or knowledge bases. The 200 topics used in the first three TREC workshops, and the relevance judgments for these topics will also be available via ftp. The topics are in the form of a formatted user need statement.

TREC-4 cont...

Queries can either be constructed automatically from this topic description, or can be manually constructed.

Two types of retrieval operations will be tested: a routing or filtering operation against new data, and an adhoc query operation against archival data. Fifty of the topics (selected from the 200 topics distributed for training) will be used by each group participating in the routing test to create formalized queries to be used for retrieval against new test data. Fifty new test topics (201-250) will be used as adhoc queries against 2 gigabytes of the training data (disks 2 and 3) plus some supplemental data (less than 250 megs).

Results from both types of queries (routing and adhoc) will be submitted to NIST as the ranked top 1000 documents retrieved for each query. Scoring techniques including traditional recall/ precision measures will be run for all systems and individual results will be returned to each participant.

New to TREC-4 will be some variations of the main tasks called "tracks". The goal of these tracks is to investigate areas tangential to the main tasks, or to investigate areas that are more focussed than the main tasks. A very brief summary of each of the 5 tracks proposed in TREC-4 is given below. The exact definition of the tracks is still being evolved by interested participants, and details of the track should be obtained from the designated contact person.

Interactive track. Investigating searching as an interactive task by examining the process as well as the outcome. Contact person: Steve Robertson (ser@is.city.ac.uk).

Multilingual track. Working with non-English test collections (250 megabytes of Spanish and 25 topics, plus possibly Chinese and/or Japanese collections). Contact person: Alan Smeaton (asmeaton@compapp.dcu.ie).

NLP track. More focussed investigation of NLP in an IR environment, emphasizing the discovery and use of phrases for TREC-4. Contact person: Tomek Strzalkowski (tomek@cs.nyu.edu).

Multiple database merging. Investigation of techniques for merging results from the various TREC subcollections (as opposed to treating the collections as a single entity). Contact person: Ellen Voorhees (ellen@learning.scr.siemens.com).

Data corruption. Examining the effects of corrupted data (such as would come from an OCR environment) by using corrupted versions of the TREC data. Contact person: Paul Kantor (kantor@zodiac.rutgers.edu).

Filtering. Evaluating routing systems on retrieving an unranked set of documents optimizing a specific effectiveness measure. Contact person: David Lewis (lewis@research.att.com).

Groups may participate in either or both of the main tasks, plus any of the tracks. There is very strong encouragement, however, to participate in the main tasks, particularly those that serve as baselines for the various tracks.

Conference Format

The conference itself will be used as a forum both for presentation of results (including failure analyses and system comparisons), and for more lengthy system presentations describing retrieval techniques used, experiments run using the data, and other issues of interest to researchers in information retrieval. As there is a limited amount of time for these presentations, the program committee will determine which groups are asked to speak and which groups will present in a poster session. Additionally some organizations may not wish to describe their proprietary algorithms, and these groups may chose to participate in a different manner (see Category C). To allow a maximum number of participants, the following three categories have been established.

Category A - Full participation. Participants will be expected to work with the full data set, and to present full details of system algorithms and various experiments run using the data, either in a talk or in a poster session.

Category B - Exploratory groups. Because small groups with novel retrieval techniques might like to participate but may have limited research resources, a category has been set up to work with only a subset of the data. This subset will consist of about 1/2 gigabyte of training data (and all training topics), and 1/4 gigabyte of test data (and all test topics). Participants in this category will be expected to follow the same schedule as category A, except with less data. New participants are encouraged to work in category B unless they have experience with such large data sets.

Category C - Evaluation only. Participants in this category will be expected to work on the full data set, submit results for common scoring and tabulation, and present their results in a poster session. They will not be expected to describe their systems in detail but will be expected to report on time and effort statistics.

Data (Test Collection)

The test collection (documents, topics, and relevance judgments) will be an extension of the collection (English only) used for the ARPA TIPSTER project. The collection was assembled from Linguistic Data Consortium text, and a signed User Agreement will be required from all participants. The documents are an assorted collection of newspapers (including the Wall Street Journal), newswires, journals, technical abstracts and email newsgroups. The test set will be of approximately the same composition as the training set, and all documents will be typical of those seen in a real-world situation (i.e. there will not be arcane vocabulary, but there may be missing pieces of text or typographical errors). The format of the documents is relatively clean and easy-to-use as is (see attachment 2). Most of the documents will consist of a text section only, with no titles or other categories. The relevance judgments against which each system's output will be scored will be made by experienced relevance assessors based on the output of all TREC participants using a pooled relevance methodology.

Response format and submission details

By Jan 16th 1995 organizations wishing to participate should respond to the call for participation by submitting a summary of their text retrieval approach, not to exceed two pages in length. The summary should include the strengths and significance of their approach to text retrieval, and highlight differences between their approach and other retrieval approaches. Groups that have

TREC-4 cont...

participated in TREC-3 need to provide only two paragraphs, one describing their methods in TREC-3 and a second describing their plans for TREC-4. In addition to the system summary, each organization should indicate in which category they wish to participate (category A, B, or C). Groups new to TREC should briefly describe their abilities to handle this large amount of data.

Please specify which main tasks and which tracks your group plans to participate in, and the person to whom correspondence should be directed. A full regular address, telephone number, and an email address should be given. EMAIL IS THE ONLY METHOD OF COMMUNICATION in TREC. The proposal should be in ascii so that it can easily be distributed to the program committee - detailed diagrams are not necessary.

All participants must be able to demonstrate their ability to work with the data collection (either the full collection or the subset). The program committee will be looking for as wide a range of text retrieval approaches as possible, and will select the best representatives of these approaches as speakers at the conference.

All responses should be submitted by Jan 16th to the Program Chair, Donna Harman (harman@magi.ncsl.nist.gov). Any questions about conference participation, response format, etc. should also be sent to the same address.

UK SALT Committee

Information Retrieval and Natural Language Processing Technology
3 & 4 January 1995
University of Sunderland

This meeting aims to:

• Promote a better understanding of information retrieval tasks and technologies amongst the speech and language communities.

• Provide a conduit for the communication experience gained from participation in US-based initiatives (MUC, TREC) to a wider British and European audience.

• Produce a strengthened pool of expertise in these important application areas for speech and language technologies.

Papers include:

Roger Moore (DRA Malvern). Topic identification in speech

Stephen Robertson (City University). Experience of TREC

Robert Gaizauskas and Takahiro Wakao (University of Sheffield), Roberto Gaigliano, Richard Morgan, Russell Collingham and the LOLITA Group (University of Durham). MUC - 6: The competitition, the controversies, and the challenges

Robert Gaizauskas, Takahiro Wakao and Yorick Wilks (University of Sheffield). Information extraction technology: A UK programme for research and evaluation

Yawei Liang and Simon French (Leeds University). Situation assessment in a teaching environment: MAP-ping the mood of student feedback based on extracting information from messages on the bulletin board

J.C. Bullock (Manchester University). Natural language generation from information encoded using semantic networks

L.J. Evett and T.G. Rose (Nottingham Trent University).Topic identification and natural language indexing for information retrieval

Karen Sparck Jones (University of Cambridge). Lessons learned from TREC and MUC

£55 Attendance, including Tea, Coffee, Lunches, Drinks Reception and Conference Dinner.
£30 Bed and Breakfast at Post House Hotel.

- John Tait (John.Tait@sunderland.ac.uk)

INET'95

The 5th Annual Conference of the Internet Society
The Internet: Towards Global Information Infrastructure
Honolulu, Hawaii, USA
27-30 June 1995

INET'95, the 5th Annual Conference of the Internet Society, focusing on worldwide issues of Internet networking, will be held on 26-30 June 1995 in Hawaii. The goal of this conference is to provide a platform that will bring together those developing and implementing Internet networks, technologies, applications, and policies worldwide for infrastructure development. The theme of INET'95 is "The Internet: Towards Global Information Infrastructure."

Since 1991, The INET conferences have become a common meeting ground for participants interested in the design, implementation, operation and use of the Internet. Global policy and economic issues, ethical concerns, and many technical issues are raised in a variety of contexts. The rapid influx of commercial and individual uses on the Internet has influenced the nature of the system and broadened its utility. The importance of the Internet and its technology to all sectors of the global economy is growing as is the social impact of access to the Internet. Internet Society encourages its members and all other interested parties to submit papers and to plan active participation in this conference.

In addition to the Network Training Workshop for Developing Countries prior to the INET Conference, the K12 Workshop, a special workshop for elementary and secondary school use of the Internet is also planned.

The conference will be organized into the following 8 tracks with example topics.

Network Technology
Broadband technology, Community networking technology, Mobility, Global networking, Next generation of Internet technology.

Network Engineering and Operation
Interoperability, Network management, Scalability, Emergency response organizations and support, Resource allocation and control, Routing and addressing.

Application Technology
Collaboration technologies, Multimedia technologies and applications, Distributed applications and environments, Networked information tools, High, low and variable bandwidth applications.

Users
Community networking services, Education and research communities, Library services and networked information, Public health and medical care, Environment,

INET'95 cont...

Entertainment, Arts and humanities.

Regional Issues
National and regional initiatives, Mission oriented networks, National and regional funding models, Empowering new users, Sociological and cultural impact, Internationalization and localization.

Policy Issues
Information infrastructure and role of governments, Policy on Internet operation, Privacy and freedom of speech, Intellectual property rights, Economy Policy.

Commercial and Business Aspects
Commercial and Business Aspects: , Commercial use of Internet, Electronic Commerce, Publication, Emerging business opportunities, Legal considerations, Global issues.

Education
Building new global learning communities, Teacher training and support models, Promoting new cultures of learning and teaching, The Internet in educational reform, Connectivity and access models, Using technology to reach learners with special needs and interest, Encouraging new partnership; industry, government, community.

In addition to the above tracks, INET'95 will encompass certain horizontal subject threads. One such thread is the World WideWeb(WWW).

13 January 1995: Extended abstract, and Tutorial proposal due
15 January 1995: Deadline for priority admission to Developing Countries Workshop

General information will be accessible on the Internet Society's WWW, Gopher, and FTP servers:

http://www.isoc.org/inet95.html
gopher: //gopher.isoc.org/11/isoc/inet95
ftp: //ftp.isoc.org/isoc/inet95

HIM'95

German Society for Computer Science Hypertext - Information Retrieval - Multimedia: HIM'95
5-7 April 1995
Information Science, University of Konstanz, Germany

More and more the requirements placed on modern information systems from the system's and the user's point of view do not allow a separate view of problems. Hence the special interest groups 'Hypertext' (4.9.1), 'Information Retrieval' (2.5.4/4.9.3) and 'Multimediate electronic documents' (4.9.2) within the German Society for Computer Science (GI), the Austrian Computer Society (VCG), the Swiss Society of Computer Scientists (SI) and the Academic Association for Information Science (HI) will hold a joint conference with focus on hypertext (HT), information retrieval (IR) and multimedia (MM).

The meeting emphasizes the following issues.

Models for HT, IR and MM. Systems' architecture and implementation. Applications of HT, IR and MM-systems. Standards for interfaces, exchange formats and query languages. Content representation in HT, IR and MM-systems. Cognitive aspects in the use of HT, IR and MM-systems. Authoring systems. Structure and administration of document sets. Distributed and open HT, IR and MM-systems. Requirements on large HT, IR and MM-systems. Database support of HT, IR and MM-systems. Cooperative, adaptive and knowledge-based methods. Interaction and user interfaces. Quality, acceptance and evaluation. HT, IR and MM-systems as means of information management

e-mail ritt@inf-wiss.uni-konstanz.de to join the e-mail conference list.

ACH-ALLC 95

Association for Computers and the Humanities / Association for Literary and Linguistic Computing
1995 Joint International Conference: ACH-ALLC 95
July 11-15, 1995
University of California, Santa Barbara, California, USA

This conference - the major forum for literary, linguistic and humanities computing - will highlight the development of new computing methodologies for research and teaching in the humanities, the development of significant new computer-based resources for humanities research, especially focusing on the issues and problems of networked access to materials, and the developing applications, evaluation, and use of traditional scientific and computing techniques in humanities disciplines.

Topics and applications are focused on the humanities disciplines, defined as broadly as possible.

Languages and literature, history, philosophy, music, art, linguistics, anthropology and archaeology, creative writing, and cultural studies. Technical proposals that focus on the cutting edge issues of the application of scientific tools and approaches to humanities disciplines; discipline-based proposals that focus on some of the more traditionally defined applications of computing inhumanities disciplines, including text encoding, hypertext, text corpora, computational lexicography, statistical models, and syntactic, semantic, stylistic and other forms of text analysis; broad library and research-based proposals that focus on significant issues of text documentation and information retrieval; and tools- focused proposals that offer innovative and substantial applications and uses for humanities-based teaching and research, throughout the academic and research worlds.

Further information from Eric Dahlin (HCF1DAHL@ucsbuxa.ucsb.edu)

ASIS 1995

Mid-Year Meeting
May 24-26
Minneapolis, Minnesota, USA

For information, contact rhill@cni.org

LISA-II

Library and Information Services in Astronomy II
May 10-12, 1995
European Southern Observatory
Garching/Munich,
Germany.

http://http.hq.eso.org/lisa-ii.html
email: lisaii@eso.org

BCS IRSG CEPIS

SDAIR '95

Fourth Annual Symposium on Document Analysis and Information Retrieval
April 24-26, 1995
Desert Inn Hotel, Las Vegas, Nevada, USA

Conference Chair - Donna Harman, National Institute of Standards and Technology.

The purpose of this symposium is to present the results of current research and to stimulate the exchange of ideas in the general field of Document Understanding. All aspects of document analysis and information retrieval are of interest, with particular emphasis on the following.

Document Analysis

Multilingual OCR, language identification, multilingual character sets, domain specific, recognition of tables and equations, recognition of maps and mechanical drawings

Information Retrieval

Full-text retrieval, retrieval from structured documents, text categorization, evaluation of IR systems, image and multimedia retrieval, language-specific influences on retrieval, text representation

Further information from Larry Spitz (Chair, Document Analysis) or David D. Lewis (Chair, IR)
c/o Information Science Research Institute
University of Nevada, Las Vegas
4505 Maryland Parkway
Box 454021
Las Vegas, NV 89154-4021, USA

ELVIRA

2nd International Conference on Electronic Library and Visual Information Research
De Montfort University,
Kents Hill, Milton Keynes
2-4 May 1995

Following the highly successful first conference in May 1994, De Montfort University will be the host of the Second International Conference on ELVIRA. The electronic library or virtual library represents one of the most important research areas in Information Sciences and Librarianship. The Conference brings together individuals from both academia and industry who are involved in the R&D of electronic library theories and systems. Due to the multidisciplinary nature of electronic library development, researchers in other related areas are also welcome to discuss their achievements in the context of the electronic library. Research students are encouraged to present their findings. The Conference will cover both technical and social-economic aspects of the electronic library, which will be addressed at both theoretical and application levels.

Possible topics include but are not limited to the following.

Electronic library theory and systems

Electronic library models, electronic library and other related concepts, electronic library systems.

Information networking

Electronic document delivery, network information retrieval (GOPHER, WAIS, WWW), protocols (Z39.50, SFQL), CWIS, high speed networks for image and multimedia communication.

Image processing, graphics and visualisation

Document image processing, digital fine-art collections, digital video, virtual reality.

Information retrieval

Neural networks and fuzzy systems in information retrieval, retrieval using parallel computers, indexing and retrieval of images and multimedia objects, information filtering, concept retrieval.

Human computer interaction

User interface of electronic library systems, ergonomics, usability tools and evaluation, user modelling, learning/educational effects.

Electronic publishing

Electronic books and journals, hypertext/hypermedia information, document architecture, document interchange standards.

Economics and management

Cost of implementing electronic libraries, economic modelling and pricing of electronic information, organisational changes.

Copyright

Intellectual property rights and electronic information, copyright management systems.

Kathryn Arnold (tel: 0908 834923, fax: 0908 834929)

PACIS

1995 Pan Pacific Conference on Information Systems
Singapore, 29 June - 2 July 1995

The theme is 'policy and strategy research in information systems: Asia Pacific perspectives'.

PACIS represents the largest gathering of information systems researchers in the Asia Pacific region. The purpose of the conference is to provide a forum for researchers, practitioners and policy-makers in the Asia Pacific region to exchange ideas on the cutting-edge adoption of information technology. In addition, there will be presentations on models of how the various nations embrace information technology infrastructure for economic success.

Topics include those pertaining to the conference theme as well as all other aspects of information system development, management, strategy and impact, including, but are not limited to the following:

Development and management of information superhighways. National information infrastructure policy and regulation. Privacy and security in transborder data flows. Role of information systems in changing economic environment. Impact of telecommunication networks and systems. IS/IT strategies in year 2000. Comparison of regional and government IS/IT policies. Implementation and management strategies for telecommuting. Multimedia and intelligent information systems. Human computer interaction for information systems. Critical success factors of business process reengineering. Enterprise systems and cooperative infor-

PACIS cont...

mation systems. Supporting object-oriented design and systems development. Groupware utilization and implementation issues. Advanced systems engineering environments. Issues in global information technology management. Quality and productivity issues in information systems Multi-social and cross-cultural studies in information systems. Client/server information systems. MIS education in US and Asia Pacific countries

- Margaret Tan (Bitnet: fbatanm@nusvm)

Workshop on social contexts of Hypermedia

16th - 17th Febraury 1995
Department of Informatics, Ume University, Sweden.

Today we are witnessing a growing number of computer applications that interlink users, databases and different kinds of media in numerous, unforeseeable ways. Often the prefix hyper, e.g. hypertext, hyperspace or hypermedia, is used in order to indicate that this development opens new dimensions of computer usage. Applications like World Wide Web have made hyper systems available to large number of users.

The field of hypermedia has been defined as concerned with the design and use of systems that support authoring, managing and navigating networks of interlinked textual and multimedia information. A large number of hypermedia systems have been constructed and various methods for their effective construction and implementation have been proposed.

In this workshop we will focus on hypermedia in use settings. What do the new possibilities of hypermedia systems mean to applications like collaborative work, decision support, education, community networks, and business administration? Often the emancipatory force of hypermedia has been stressed. On the other hand, in many contexts, it has to be proven that hypermedia can be used to a practical purpose.

Other questions are the way hypermedia affects the nature of knowledge. Does the break up and non linearity of information in hypermedia systems mean more fragmented knowledge?

What requirements must be put on the computer human interface and the organisation of information in order to navigate through the information without getting lost?

The purpose of this workshop is to discuss the theoretical basis for the study of various social aspects of the use of hypermedia systems and its implications for the design of such systems.

Some questions related to this purpose could be: What interdisciplinary connections could be made for example to psychology, theory of literature and philosophy? What kind of representations of information in hyperspace are helpful for users? How can different retrieval mechanisms (search, navigation) and styles of interaction (conversational, direct manipulation) best be integrated to meet users' task needs? Which factors are of importance for the design of the interface for computer human interaction? How does the user interface model affect the perception of the knowledge base (e.g. in terms of coverage, interrelations, etc.)? How can hypersystems be helpful in various work situations? How can they be used for the benefit of disabled persons? How can hyperspaces best be shared? Issues of collaboration and concealment? Personalisation for individuals and groups - customisation by users versus automatic adaptivity?

Invited speakers:

Peter Bogh Andersen, Department of Information and Media Science, Aarhus University, Denmark.

Mark Chignell, Department of Industrial Engineering, University of Toronto, Canada.

Kaj Gronbak, Computer Science Department, Aarhus university, Denmark.

Norbert Streitz, Integrated Publication and Information Systems Institute, German National Research Center for Computer Science, Darmstadt, Germany.

Randall H. Trigg, Xerox Palo Alto Research Center, USA.

Mika Tuomola, Continuing Education Centre, Theatre Academy, Helsinki, Finland.

John Waterworth, Department of Informatics, Ume University, Sweden.

The workshop fee is 3000 Swedish Crowns (about US $400). If you want to stay in a a hotel in the center of Ume you will have to pay 3250 Swedish Crowns (about US $435).

Futher details from Kenneth Nilsson (Kenneth.Nilsson@informatik.umu.se) or from http://www.informatik.umu.se

A wee bit of history

Looking back through the few past issues of the newsletter I have, I wondered if anyone else might have backcopies that they'd be willing to donate to the 'archives', i.e. me. If so, I'd be very interested to hear from you.

The first copy I've got is issue 30 from May 1986. Back then the newsletter had its own ISSN (probably still does) and was called 'IR Information Retrieval', simple and accurate but somewhat lacking in style. The editor isn't listed but I would guess from the writing style that it was John Lindsay.

For a few years, from 1986, it included the results of a Selective Dissemination of Information (SDI) search on the Inspec database that the group used to do (at a certain cost) for the benefit of all. This has since been dropped on the grounds that the expense outweighed the interest. One way of making it more useful might be to carry out a wider-ranging search at a greater cost which could be partially covered by sponsorship from interested members. Any comments?

A couple of years on, Winter 1988 included a financial report. I've got a copy of the 93/94 financial report with me and it's 8 pages long. Would members want to see this in the newsletter?

Mark Sanderson took over in Summer 1992 and since then the newsletter has been altogether less dry and basically an entertaining as well as an informative read. Mark's milestones have included the first photograph (an Apple Newton) and the first mention of a member of the British IR Community throwing up (on a fairground ride at SIGIR 92, Copenhagen). I hope to include more photos (but not more vomitting) in future editions.

- Mark Magennis, Editor

The thoughts of Chairman Mark

The first in a series of Informer interviews with key members of the British IR community. This time it's the turn of Dr Mark Dunlop.

Dr Dunlop, what is your position in the BCS IRSG?

Chair (usually in the corner).

What is your job title?

Ah, good question, now is it 'academic' or 'lecturer'? Never really sure.

What projects or ideas are you currently working on or planning to work on in the future?

I'm involved with the EU working group MIRO looking at multimedia indexing, upcoming proposals to the EU on multimedia object content analysis (MOCA) and on the affects of IR on users' primary tasks. I am still playing around with the combination of hypermedia and free text retrieval (my thesis was in this area). The latest thing is an interesting discovery about how bad negative feedback is.

What other areas of IR are you interested in?

Phew, anything on the user-oriented side really. Evaluation is getting more interesting, user interfaces always have been. Just don't worry me with your logics and mathematics.

What areas outside of IR are you active or interested in?
Human computer interaction is my main computing interest outside IR. Outside computing my most active interest is fighting against Glasgow rain to keep rust at bay on my 18 year old MG.

What do you wear under your kilt?

A big happy smile!

What was the greatest moment of your career (or what will it be when it finally happens)?

When a class representative said my lectures were "inspirational", and yes I am going to keep quoting her.

What are your main interests outside work?

Och, carburetter balancing and rust-prevention. Also less anorachy things like movies; food (cooking and eating - you've got to try my chocolate ice cream, the trick to soft scoop home made is litres of alcohol); curling (no it is not an old man's game, you try sweeping in front of a lump of granite while running on ice); whisky (any kind - I'm not proud, but preferably a nice peaty malt); being with my wife (ahhhh, how sweet).

Why are there so many Marks from Glasgow on the committee of the IRSG?

Nice weather we're having.

Predict the future of IR - what do you think will happen in the next 20 years?

Hopefully I will be able to stop saying "it's like databases only less structured" as an opening line. I think we might just be on a breakthrough point to commercial use of IR techniques. I also see it a bit like AI (whoops he mentioned the A acronym), I think we have to keep looking toward the Holy Grail of perfect retrieval but in the meanwhile IR will be used in lots of little places and interfaces will make the mistakes less critical.

What future directions can you imagine for the BCS IRSG?

There have been some very promising moves lately to make the group more European, for example the last Colloquium was greatly helped by collegues from Eire, Holland, France.... This looks good. I also hope it continues to be a strong focus and social meeting for the UK groups. The emphasis on young cutting edge research - done by the teaching and administratively free (e.g the young) looks like strengthening and making the colloquia exciting beds for discussing new fresh ideas.

What do you think is the most neglected area of IR?

The poor user - let's remember her and try to help her cope with the mistakes of our systems.

How would you improve the BCS IRSG?

Try to make it bigger and better but still keep the small feel to the colloquia so we don't loose the young researchers feel. A better name would also be good, maybe you could run a competition - maybe a prize, let me think.... "The British Computer Society Information Retrieval Specialist Group" does lack a certain fizz.

Crack an IR-related joke (apart from the one about the chair being in the corner).

A formal methods researcher, a C++ programmer and an IR researcher were driving home after some skiing in the Alps. The brakes on the car failed and, after skidding round many bends just managing to keep the car from flying into the gully beside them, they ground to a halt in a sand bed. After much wiping of brows, swearing and deep breathing they got out the car and tried to decide what to do. The formal methodist, being sound in thought, decided the best option would be to wait for a passing car and ask them to call a garage from the bottom of the hill. The C++ hacker, being a typical computing scientist, decided that brakes aren't that complicated and he would have a look. The IR developer thought for a while, decided it might just be a flook and said "let's try that again, it might work better this time".

Begging section

In order for the Informer to 'inform' you of what's going on in IR I need 'information' to put in it. That's where you come in. I want you to send me items of news or anything else you come across that you think might be of interest to our members. This could concern the latest research, commercial products, political or social developments that affect our research and its application, forthcoming events, whatever you think we all ought to know. Many of our members won't have a handle on your own particular areas of interest so if you'd like to publicise it to a wider audience then why not consider writing an article for the Informer? I would also welcome any titbits, diary entries, opinions, observations, jokes, cartoons, malicious gossip, unsubstantiated rumours, etc. Photos or artwork would be especially appreciated.

- Mark Magennis, Editor