Information retrieval and digital libraries: lessons of research

Karen Spärck Jones
Computer Laboratory, University of Cambridge
William Gates Building, JJ Thomson Avenue, Cambridge CB3 0FD, UK
sparckjones@cl.cam.ac.uk

This paper in its final form is in the Proceedings of the *International Workshop on Research Issues in Digital Libraries (IWRIDL 2006)*, Kolkata 2006, ACM, 2007.

Abstract

This paper reviews lessons from the history of information retrieval research, with particular emphasis on recent developments. These have demonstrated the value of statistical techniques for retrieval, and have also shown that they have an important, though not exclusive, part to play in other information processing tasks, like question asnwering and summarising. The heterogeneous materials that digital libraries are expected to cover, their scale, and their changing composition, imply that statistical methods, which are general-purpose and very flexible, have significant potential value for the digital libraries of the future.

1 Introduction

The "digital library" is an attractive idea. But what does it mean, and how might we get there? Some views are comprehensive, and ambitious (e.g. Borgman 2000). Some believe that digital libraries, while unlike conventional libraries in many respects, will be like traditional libraries in being subject to quality control, both in admitting material and in describing (cataloguing) it. Others believe we have digital libraries already, thanks to Web engines, where the conventional notions of quality control certainly do not apply. This difference reflects more general differences between the library and computing communities, though recent developments in relation to the Semantic Web indicate that some in the computing community believe strongly in descriptive quality control. But paradoxically, those advocating the Semantic Web may, judging from the library community's experience over many years, be being unrealistic about the extent to which information can be fully and unambiguously described and manipulated.

In the rest of this paper I will explore the relationship between information retrieval (IR) and related research and digital libraries.

IR research, i.e. research into automated indexing and retrieval, has been done for fifty years, and has become increasingly solid. But its impact on operational library and information systems has been slow and uneven. We should therefore ask what IR research

achievements so far have been, and what implications these have for digital libraries now. Do digital libraries, on rational interpretations of what "digital library" may mean, offer new scope for the IR research methods that have already been developed, or is there a need for new methods?

2 A brief history of IR research

In this summary I will treat information, document, and text, as the same.

The starting point in automated IR research focused on the core tasks of indexing and searching (more strictly, matching). The retrieval context that motivated the collection of documents for the file, and the nature of the users and their needs, was taken as implicit in the given set of documents, the information requests as submitted, and the relevance assessments that users supplied for the search output. The presumption was that the primary requirement was for search effectiveness, i.e. returning relevant documents. Research progress on indexing and matching methods was (to be) promoted through an appropriate system evaluation methodology, specifying detailed test designs and applying numerical performance measures (for which the relevance assessments were gathered).

The findings that emerged from this research were first, that indexing and searching could be effectively done with document descriptions that were derived from the document texts themselves, rather than assigned to them; second, that these descriptions were best grounded in distributional data about words or other simple linguistic units, and third, that statistical techniques, when applied to this distributional data in combined indexing and searching operations, delivered retrieval systems that actually worked. Specifically they worked not only as well as the previously existing manual systems, but better.

These points are now taken for granted by those engaged in IR research, and by some others as described later. But it is important to examine them in more detail for their implications for digital libraries, especially when digital libraries are regarded as natural extensions of familiar, non-digital libraries, particularly in terms of indexing and searching, even if the materials collected can be novel in form or content.

Early IR research was not merely focused on the core tasks of indexing and searching. The assumption was that the goal was to automate existing manual indexing and searching strategies. However quite novel strategies emerged from this effort which were important not only for retrieval in the narrow sense, but linked this with other information management tasks, like summarising, in a way that has become increasingly important.

Thus Luhn's early work, from 1957 onwards (see Schultz 1968), was intended to provide support for human indexers. The statistical information he computed about the associations between words in texts, and about relative word weights, was initially envisaged as supplying support for the human indexer in choosing important concepts to describe documents. However Luhn saw that the same general ideas could be extended to other information management tasks, in particular to summarising by statistically-weighted sentence extraction. We can see now that this work opened a Pandora's box in automating processes for reaching and using linguistically expressed information.

It became evident in particular that indexing and searching could be fully, not merely partially, automated along these lines. Three lines of work contributed to promote this.

First, the development of an underpinning theory for the statistical approach to retrieval, notably by Maron and Kuhns (1961). This provided an interpretation for the key notion of

probability of relevance, using statistical term weights and term associations, search output ranking, and statistical relations between documents in iterative feedback.

The second contribution was Cleverdon's development, from 1960 onwards, of detailed methods for index and search testing (Cleverdon 1967)). These supported systematic, controlled experiments based on a specification of the factors affecting system performance, of the devices intended to achieve particular performance effects, and of concrete performance measures. This work established the evaluation paradigm, centred on precision and recall, that we still use. Cleverdon himself did not then work with automated systems.

The third contribution was Salton's development of computational systems, from 1962 onwards (Salton 1968, 1971). His systems had a text-statistical base, and were tested in decompositional laboratory evaluations. Drawing the previous work together and building carefully on it, Salton showed that the best retrieval performance was obtained with the text of abstracts (extended full text was not generally available), with word stems, with statistical weights, with statistically-based matching scores giving a ranked output, and with relevance feedback. Performance could benefit a little from adding a manual thesaurus, but even without this, good, competitive systems could be built. There was no justification for the controlled indexing and boolean searching of the conventional library world.

However all these early tests were with very small data sets, and performance varied with different collections.

IR research in the 1970s and 1980s built on this initial work, trying to consolidate it. It thus explored more, and different, formal models for retrieval systems, like inference networks (Croft 2000); conducted tests with bigger and more varied test collections, e.g. ones with 11,500 rather than 1000 abstract texts, and investigated other performance measures, like document cutoff. Overall this research worked on an ever-wider range of issues, and in ever-finer grain, investigating the effects both of changes in environment variable values, like document types, and of changes to system parameters, like methods of clustering documents. The results were both a better understanding of retrieval itself, for instance of the effects of uncertainty (e.g. about what the user's request means) as a constraint on performance, and more evidence for good methods (e.g. how pervasively helpful statistics can be at all points in system processing).

Thus the research mainstream, devoted to the core indexing and searching processes, established that simple natural-language indexing is good, that statistical term weighting is good, that relevance feedback is good, and that ranked output is good. It also established that clustering, either for terms or documents, is not unequivocally good, and that sophisticated grammatical analysis is not good. These findings in themselves contradicted received library service wisdom. The research thus established, in specific contradiction to received wisdom, that in general there was no gain from using a (manual) thesaurus or subject headings, and that boolean matching is inferior to best-match ranked.

But even after these two decades, this research work had limitations: the tests were still small scale, they were remote from users and their context, and there was no real interactive search. Thus the outcome of thirty years research was a black box in an opaque mathematical packaging, in a world apart from operational systems.

Why was this?

The operational library service world had been automating: first, for existing libraries, in cataloguing (as with MARC and OCLC); then for other types of bibliographic resource, like abstracting services, in automating databases, searching, and retrieval (as with INSPEC and MEDLARS); then by going online for these services (as with DIALOG, ORBIT, ESA,

MEDLINE); and also by exploiting full text in services (as with LEXIS and STAIRS). (The only wholly novel system in this period, and exception to the others, was SCI.)

In these systems, whether for catalogue files or for more comprehensive ones with abstracts etc, conventional assumptions about the system core were entrenched, i.e. that controlled language indexing and boolean matching were required. At the same time, as these earlier large automated systems were developed, they had to address many concerns that mattered for users, like document delivery, file coverage, interface design, and multiple languages, which researchers could or did ignore. There was nevertheless some convergence with research in a growing use of natural language indexing, at least alongside controlled, and of full text. There were also, by the early 1990s, some operational systems using ideas from IR research (see Tenopir and Cahn 1994).

3 The present research state

The 1990s saw a revolution in IR research. First, there were two major changes in the research environment. There were radical developments in information technology (IT) in general; and there were significant developments in natural language processing (NLP) technology, and specifically in such processing aimed at information access and management tasks like summarising, which may be called natural language information processing (NLIP).

The IT developments are familiar, but may be briefly listed for their importance for library and information systems. Thus machines have become enormously more powerful and much more comprehensively connected; they have been flooded with bulk "stuff", of every kind including multimedia material as well as masses of text. More importantly, the Web has arrived.

NLIP research has advanced sufficiently for there to be non-trivial systems for tasks like summarising or question answering. There are also well-founded portable tools for particular processes, like parsing, and there are general techniques, like Hidden Markov Modelling, that are applicable in many contexts. But the most important recent development in NLIP as a whole has been in the growth of evaluation programmes.

The combination of these IT and NLIP developments, and in particular the joint appearance of the Web and the rise of evaluation programmes, has had major effects both on IR research and on the relations between IR research and the 'real' world.

Thus the Web contains vast, mixed data, not just the 'proper papers' on which bibliographic information services and research alike have so long concentrated. It has a huge and varied clientele, not just the sort of 'serious user' familiar from e.g. scientific search services. It has an enormous spread of assorted search types, and not just the 'regular topics' that professional services, e.g. for medical researchers, have taken for granted.

The search engines the Web has stimulated have, as natural responses to what they have to deal with, been thoroughly eclectic. In particular they exploit a far wider range of factors and devices than either conventional services or researchers have done. At the same time, they have taken some key devices from mainstream IR research, and have ignored or abandoned many conventional ones. Thus the engine builders have adopted statistical techniques as both conceptually appropriate, especially for large-scale applications, and practically convenient; and they have recognised that classical indexing languages cannot be built for or applied to heterogeneous Web worlds.

The evaluation programmes, sponsored primarily by US agencies such as (D)ARPA, NIST

150 requests, 370 K documents, full text

precision at rank 10

	10 terms	4 terms
unweighted terms	.11	.15
basic weighted	.52	.47
relevance weighted, expanded	.61	.51
assumed relevant	.57	.46

Figure 1: Retrieval performance with statistical methods (from Sparck Jones, Walker and Robertson 2000)

and ARDA, but also Japanese and European agencies through the NTCIR and CLEF programmes, have addressed a range of NLIP tasks including speech recognition, information extraction, text and other retrieval, and summarising. The Text REtrieval Conferences (TRECs) (TREC 2006; Voorhees and Harman 2005), that have now been running for fifteen years have been particularly important for their scale and scope, and have had a wide influence not only on retrieval but on other task areas, for several reasons. They have been influential through their choice of tasks, the methodologies they have established, the many researchers they have involved, and the results they have obtained.

The TRECs have engaged in systematic, controlled tests, over many cycles in which conditions have been carefully varied. For retrieval they have used very large collections and, with many participants, have tested many systems. These experiments have thus covered a very rich range of comparisons, and delivered very solid results.

3.1 Text retrieval

The most notable outcome is that for classic topic searching in retrieval, i.e. for documents about X, these evaluations have confirmed previous research findings. The value of the statistical techniques developed in IR research over the previous decades can be illustrated with some TREC data results taken from Sparck Jones, Walker and Robertson (2000). Thus Figure 1 shows that retrieval performance improves markedly when search terms are statistically weighted, and even more when relevance feedback information is used to expand requests with statistically preferred terms. There are performance gains even with very short requests. Moreover with longer requests, even assuming (rather than actually knowing) that the best matching documents on a first search pass are relevant, and expanding queries accordingly, can improve performance, though this may not (as here) hold for very short queries.

3.2 Enlarging the retrieval envelope

TREC retrieval testing has also been extended from the classic English monolingual case. TREC experiments have shown that the same good, statistical indexing and searching methods work with very different languages, like Chinese; they also work for multi-lingual files where queries need translation. Again, they can be applied to newer, non-standard document types like many web pages, and to index keys like links and URLs which are not ordinary

50 requests, 21 K news stories in 28K items

mean average precision

	11 words		3 words	
	HUM	SR	HUM	SR
known boundaries -				
basic weighted	.38	.35	.43	.40
assumed relevant	.43	.37	.47	.44

Figure 2: Retrieval on transcribed spoken news stories (taken from Sparck Jones et al. 2001)

language objects. These statistical methods have also been shown to be effective for transcribed speech, which is very noisy text, though there are many problems with content-based image retrieval when any associated language data is lacking. Figure 2 shows performance for retrieval on automatically transcribed speech (SR) when compared with correctly (i.e. manually) transcribed versions (HUM). Actual retrieval performance can be quite competitive even where transcription word error rates are more than 10%.

TREC has also promoted work on tasks such as text routing and filtering, that are closely related to one-shot text searching but have their own distinctive properties. Thus filtering requires yes/no decisions for output, not ranking. Here again, statistical techniques are very effective.

3.3 Further enlarging the envelope - question answering

But going beyond these somewhat incremental developments within the general area of text retrieval, TREC has initiated a new and important stream of research on a significantly more challenging task, namely question answering (QA). In conventional retrieval, requests for information may be submitted in the form of questions, but they are always treated as simple topic specifications, seeking documents about the topic. The TREC QA track has addressed the much more difficult task of answering specific questions, like "Where is the Taj Mahal?", where the answers take the form of exact text snippets extracted from the file, for example in this case "Agra, India". In general, answering questions may require the generation of new linguistic output, but for many purposes appropriate quotation from an existing text source may be adequate and hence require a less comprehensive overall system.

The QA tests have gradually extended from answering simple 'factoid' questions, like the example just given, to providing list answers, like 'What are five of Dickens' novels?' which may have to be assembled from more than one source, and to dealing with definition-type questions like 'Who was X?', where an appropriate response would assemble several items of information about the person.

Just as with other NLIP tasks, including retrieval, question answering is not a single, well-defined task. It has many variations depending on contextual factors, notably the user's encompassing purpose as well as the characteristics of the data where answers are sought. QA techniques correspondingly range from primarily statistical NLP to primarily symbolic NLP. Statistical passage (not snippet) extraction, following retrieval models, is crude but may be helpful, though the user has still to find the answer within the returned passage. Combining very lightweight NLP methods with statistical ones, e.g. in looking for answer patterns and

```
Where did Dr King give his speech in Washington?

[candidate passage containing answer, abbreviated]
... Dr ... King delivered his ‘‘I have a dream’’
speech at the Lincoln Memorial, ...

     ==> Lincoln Memorial      [Yang & Chua 2002]
```

Figure 3: Question answering test performance example (reduced from Yang and Chua 2002)

exploiting frequency data taken from large corpora like the Web, can be suprisingly effective. However much more power is obtained (at more system development cost) by exploiting heavyweight NLP methods involving deep text interpretation and connective inference, with support from statistics e.g. about word-cooccurrence relationships.

The important lesson learnt from the QA research of the last decade is that while effective exact question answering cannot generally be done using purely statistical techniques, so symbolic language processing is required, statistical data and methods play an important part. Many systems combine the two types of technique, with statistical information about word (or word string) behaviour in large text files acting as a substitute for the kind of elaborate and explicit characterisation of the world (as envisaged in the Semantic Web), that cannot in practice be provided for vast, completely general, and constantly changing text collections like the existing Web. Many systems also make use of statistical retrieval techniques to identify passages for more detailed candidate answer analysis.

Fig 3 illustrates an output for one system combining statistical and symbolic NLIP techniques, where data about relative pattern frequencies over words and word combinations are used in conjunction with structural characterisations of sentences (Yang and Chua 2002). Progress examples like this for a difficult task are encouraging, but current QA systems do not work anything like as well consistently.

The TREC and ARDA AQUAINT programmes have contributed enormously, through the systematic evaluations they have sponsored, to promote QA system development. It is, however, extremely difficult to evaluate question answering when dealing with complex questions like definition ones, since there are in general no unequivocal correct sets of ‘facts’ that define people or things. Evaluation has also to allow for variant surface formulations of the same underlying content. All of the evaluations done so far have had to compromise, by accepting some artificial constraints on the task interpretation, so as to be able to deliver measured results. The realities of question answering are that all answers have to be interpreted and assessed by human users, and that this interpretation and assessment is subject both to what the users know and to what the user’s motivation is on each particular occasion.

Figure 4 shows this with a very simple example where TREC-type QA might return any of the listed snippets. They are progressively further from the first ‘correct’, exact answer, but this does not imply that any of them, even the last, is of no value to the user since something like the same answer can be seen in, or inferred from, the system return. Moreover, even with the first response, the user has to decide whether they think it is correct. As these examples imply, context-based NLIP task evaluation can be extremely difficult, and needs developing and specifying with great care (Sparck Jones and Galliers 1996).

```
What is the longest river in the United States?

   the Mississippi
   the mississippi River

?  2,348 Mississippi

?  At 2,348 miles, the Mississippi River is the
     longest river in the US.

?  The Mississipi stretches from Minnesota to
     Louisiana.
```

Figure 4: Alternative answers to a question

3.4 Pushing out of the envelope - summarising

The most radical push on the retrieval envelope has been through the DUC programme on automatic summarising (see DUC 2006). In the same way that QA touches on retrieval by sharing the notion of information seeking, summarising is linked to retrieval by sharing the indexing notion of brief information characterisation. Equally, developments in summarising research in the last decade have demonstrated that statistical methods have an important part to play for this task too, especially where summarising is focused on the selection of key material from source texts. As with question answering, the techniques studied range from the purely statistical to combinations of statistical and symbolic methods. Using lexical frequency data to extract important source sentences, or multi-sentence passages, is crude but may be useful. Strategies that combine such frequency data with lightweight parsing to identify key entity references and resolve anaphors may give more discriminating extraction. Finally, more comprehensive parsing can be used to identify both major rather than subordinate information within individual sentences and information shared by many sentences when summaries are produced for multiple documents rather than single sources. In each case the selected sentence components form the base for output text sentence generation.

Figure 5 illustrates output for a system producing summaries over sets of documents in response to complex topic requirements (Lacatusu et al. 2005). Here symbolic parsing is used to decompose the topic into its components which are analysed using both statistical frequency information and parsing to identify key component concepts. These concept sets are used to score source sentences as candidate material for the final summary, where the sentence parses also make it possible to reject duplicated information. Similar combined methods are used in the operational Newsblaster system developed at Columbia University (Newsblaster 2006).

Again, the evaluation programme that has encouraged summarising system development has also shown how difficult it is to evaluate system performance for complex NLIP tasks. There are certainly no single correct summaries for a given source, so evaluation based on comparing system outputs against model summaries, whether for common extracted sentences or for shared content 'nuggets', is not necessarily a good guide to summary values for users with particular purposes in particular context. However evaluating actual contextual effectiveness, whether for QA or summarising, is difficult and expensive. Retrieval testing

```
Topic:
What are new technologies for producing steel?
What are the technical and cost benefits of
these new technologies over older methods ...

Topic-responsive summary, abbreviated:
Nucor, which has pioneered the use of a cost-
effective new steel-making technology called
thin-slab casting is one of the lowest cost
manufacturers of steel in the world. ... thin
slab ... The success of Nucor ... whether large,
labour-intensive plants can survive ... Brussels
says ... cut 30M tonnes of crude steel ...
                        [Lacatusu et al 2005]
```

Figure 5: Topic-responsive multi-document summarising example (reduced from Lacatusu et al. 2005)

```
Stockbrokers are reporting a 'spectacular' increase
in online trading as private investors storm back
into the market after five successive quarters of
declining business.

?  Private traders storm back to markets.
?  Large increase in online trading.
?  Spectacular increase in private investor trading.
?  Online private traders back after long break.
```

Figure 6: Alternative summaries for a text

has always referred to purpose and context through the use of relevance assessments, but this has been in a limited, though still expensive, way. There are no analogues of relevance assessments for question answering or, especially, summarising, that can be taken as wholly respectable substitutes for full system evaluation in live user situations.

Just to emphasise this point, Figure 6 shows alternative, quite different but equally plausible summaries of a short text. Their relative merit can only be established by their value in some larger task context.

Even so, it is possible to claim some progress in building NLIP systems for a range of tasks over the last decade, i.e. progress in building systems that are performing better both intuitively and against reasonable though not ideal criteria. Moreover what has been learnt from building what are essentially general-purpose systems appears to provide a valuable platform for customisation to particular applications.

3.5 Gains from statistics

This advance is particularly clear in relation to statistical methods. Thus these have not only been proven valuable in themselves. They are also intrinsically tailored, i.e. they are based on whatever material is necessarily supplied for an application. From this point of view, one

of the major steps in the last decade has been the appearance of a unifying model, so-called Language Modelling, which can be applied to a range of NLIP tasks, and can be very finely tuned to the form of the datasets that are associated with a type of task and to the dataset content for a particular application (see Croft and Lafferty 2003).

Language Modelling embodies what may be described as 'the ngram revolution', where natural language meaning is only implicit in the identification and use of character/word string behaviour, but is nonetheless effective for being implicit rather than explicit.

The essential idea in Language Modelling is that given a corpus of paired discourses, A and B, correlations can be established between features of A and features of B (the features being e.g. word sequences, or sets), so that for a new A, a new B can be derived. In speech transcription (the original locus), A is a sound stream and B is text; in translation A is source text and B is target text; in summarising A is a source document and B is a summary; and in retrieval A is a request and B is a relevant document. This very general technique works extremely well on some tasks, notably transcription, is effective for retrieval, and has been interestingly illustrated for translation and summarising. With the large training corpora now available, Language Modelling is an active area of research.

3.6 Some observations on the present state

In reviewing all this development, the following observations can be made about the relations between retrieval (or information management) research, search engines, and libraries.

In libraries, automation preceded innovation (as with OCLC). Innovation was forced by computing researchers (e.g. through the Web, from AltaVista onwards). Indeed many retrieval researchers have been in computing departments, not library schools. Libraries have been slow to take up research ideas.

This has partly been for good reasons: the research ideas have been unproven, disruptive, and costly; and other factors than those on which research has concentrated have dominated perceived library performance. But it has also been partly for bad reasons: general inertia and the 'not invented here' syndrome. Professional library hostility to upstart computing researchers is either a good or a bad reason, depending on viewpoint.

Computing researchers on retrieval and related tasks have come, unlike librarians, without any intellectual baggage. This has had good effects, in allowing rapid action on new ideas and in encouraging boundary crossing. But it has also had bad effects, notably in ignoring real library experience and in reinventing wheels, as currently illustrated by the Semantic Web movement.

Overall, however, the advances made by researchers in information retrieval and allied tasks are extremely important, and the digital library movement needs to exploit them. Certainly, frequently-repeated statements that are made about the comprehensive scope that digital libraries are meant to have in the IT-based future imply that recent and current experience in NLIP research needs to be taken on board.

The particular findings I believe should be exploited are summarised in the next section.

4 Research implications for digital libraries

My analysis assumes that the important infrastructure issues, for example about document formats, interoperability and the like, have been addessed. My focus is also on what is still the main vehicle for conveying information, namely natural language text. I am excluding here

all types of non-language multi-media e.g. images without associated language data, though these are clearly extremely important. Except in special, normally restricted, circumstances, material without associated language data is extremely difficult to access, as in pure image retrieval. However where, as is often the case, there is associated language data, the methods condidered here naturally apply.

It also goes without saying that general, operational developments in computing, and of the Web, should be exploited: digital libraries should welcome new supports for information management and use, like mobile access, new information objects of unfamiliar kinds like lecture slides, and new information cues, like URL links. Digital libraries should also, naturally, make use of any forms of information management that the Semantic Web movement delivers, though these are more likely in practice to be specific tools for particular domains than universal tools. Indeed observation of the Web suggests that the idea that effective digital libraries require well-defined and detailed ontologies is fundamentally mistaken. Classical library protocols, that involve human quality control both on material admitted to a collection and to its indexing will not in general work for future digital libraries, outside restricted contexts. The appropriate response to managing the digital libraries of the future is therefore not to seek to impose a prior order on the library's content, but to make such order as there actually is within it to emerge.

Thus my basic message is that those seeking to develop digital libraries should implement the findings of retrieval research, i.e. should adopt statistical, text-based technques, and should import, as appropriate, other technologies like speech processing and natural language processing.

This may seem obvious and uncontroversial. However my specific recommendation is the much more pointed one, namely to apply, systematically, the general retrieval research lesson: this is to use statistical data as far as you can (and seek always to go further). There is bulk language data for the asking, and more all the time; and there are general, available methods for processing it. These may be called pattern matching, classification/clustering, or machine learning, but they are all methods for 'finding like things' which is what information search is all about.

Statistical methods are very good for some NLIP tasks, for example document retrieval, and speech recognition (i.e. transcription). They are also quite adequate for 'down-market' versions of some tasks, for example indicative summarising and selective text extraction, i.e. for forms of NLIP tasks where the context and purpose are not very demanding. Finally, they are helpful for, indeed may be crucial contributors to, complex NLIP tasks, with a role in supporting particular subtasks or at particular processing stages, as for example in question answering and multi-text summarising.

Statistical methods are further valuable first, because they promote multi-task integration: their generality encourages a common perspective on tasks that may in fact share some features even if they differ in others, as with retrieval and summarising; and second, because their very simplicity encourages easy trials, for example of ways of giving query-oriented summaries for retrieval system outputs. Web engines, like Google, already show query-oriented snippets from retrieved pages. The underlying retrieval mechanisms are essentially statistical, though the snippets appear to be selected on a much cruder basis. Research is already beginning, however, on improving the quality of such summaries by using statistical and/or simple symbolic techniques.

The essential point about using statistical methods for NLIP tasks, i.e. to process language material, is that model and tasks clearly fit one another. Statistical methods work through

redundancy: they identify patterns in noise. All use of language has redundancy: words and word strings are ambiguous, and the ambiguity is only resolved by patterned reptition. This convergence clearly implies that statistical strategies are the right basic tools for large-scale information management.

The arguments advanced in this paper may seem hostile to the principles and practice of one of the father figures of library science, S.R. Ranganathan. Ranganathan's work was primarily done before automation introduced new possibilities for characterising and finding information. But this is not the significant point. While "The Colon Classification" (Ranganathan 1965) might appear to imply a universal classification of the traditional kind, it was the method, along with a few very high-level organising concepts, that was universal. The approach could thus be applied to faceted classification schemes for specialised literatures (Vickery 1966). Moreover the whole aim was to be responsive to an actual literature, reinterpreting the organising concepts (categories, facets) for those contexts. Modern statistical techniques for describing and manipulating textual information are specifically intended for derivative, i.e. responsive, information description. They can thus be justified as adopting the same essential approach to indexing and retrieval, albeit in an implicit rather than explicit (and perhaps also less schematically tidy) way, as Ranganathan did.

References

Borgman, C, L. *From Gutenberg to the global information structure*, Cambridge MA: MIT Press, 2000.

Cleverdon, C.W. 'The Cranfield tests on index language devices', *Aslib Proceedings*, 19, 1967, 173-192. Reprinted in Sparck Jones and Willett (1997).

Croft, W.B. (Ed.) *Advances in information retrieval*, Boston MA: Kluwer, 2000.

Croft, W.B. and Lafferty, J. (Eds.) *Language modelling for information retrieval*, Dordrecht: Kluwer, 2003.

DUC: see http://duc.nist.gov (visited 2006)

Lacatusu, F. et al. 'Lite-GISTexter at DUC 2005', *Document Understanding Workshop (DUC) 2005*, Proceedings, 2005, 88-94; via http://duc.nist.gov.

Maron, M.E. and Kuhns, J.L. 'On relevance, probabilistic indexing and information retrieval', *Journal of the ACM*, 7, 1960, 216-244. Reprinted in Sparck Jones and Willett (1997).

Newsblaster: see newsblaster.cs.columbia.edu (visited 2006)

Ranganathan, S.R. *The Colon Classification*, Vol. 4, Rutgers Series on Systems for the Intellectual Organisation of Information, Graduate School of Librray Service, Rutgers University, 1965.

Salton, G. *Automatic information organisation and retrieval*, New York NY: McGraw-Hill, 1968.

Salton, G. (Ed.) *The SMART retrieval system*, Englewood Cliffs NJ: Prentice-Hall, 1971.

Schultz, C.K. (Ed.) *H.P. Luhn: Pioneer of information science*, New York NY: Spartan Books, 1968.

Sparck Jones, K. and Galliers, J.R. *Evaluating natural language processing systems*, Berlin: Springer, 1996.

Sparck Jones, K. and Willett, P.H. (Eds.) *Readings in information retrieval*, San Francisco CA: Morgan Kaufmann, 1997.

Sparck Jones, K., Walker, S. and Robertson, S.E. 'A probabilistic model of information retrieval: development and comparative experiments. Parts 1 and 2', *Information Processing and Management*, 36 (6), 2000, 779-808 and 809-840.

Sparck Jones, K. et al. *The Cambridge Multimedia Document Retrieval (MDR) project: Summary of experiments*, Technical Report 517, Computer Laboratory, University of Cambridge, 2001.

Tenopir, C. and Cahn, P. 'TARGET and FREESTYLE: DIALOG and Mead join the relevance ranks', *Online*, 18, 1994, 31-47. Reprinted in Sparck Jones and Willett (1997).

TREC: see http: //trec.nist.gov (visited 2006).

Vickery, B.C. *Faceted classification schemes*, Vol. 5, Rutgers Series on Systems for the Intellectual Organisation of Information, Graduate School of Library Service, Rutgers University, 1966.

Voorhees, E.M. and Harman, D.K. (Eds.) *TREC: Experiment and evaluation in information retrieval*, Cambridge MA: MIT Press, 2005.

Yang, H. and Chua, T.-S. 'The integration of lexical knowledge and external resources for question answering', *The Eleventh Text REtrieval Conference: TREC 2002*, Special Publication 500-251, National Institute of Standards and Technology, Gaithersburg MD, 2002, 486-491.