Information retrieval: how far will *really* simple methods take you?

Karen Sparck Jones
Computer Laboratory, University of Cambridge
New Museums Site, Corn Exchange STree
Cambridge CB2 3QG, England
sparckjones@cl.cam.ac.uk

Prepared for
TWLT 14, Language Technology in Multimedia Information Retrieval, 1998

Abstract

This paper considers document (text) retrieval techniques and their relations with natural language processing (NLP). It summarises retrieval requirements, illustrates current approaches to retrieval using a probabilistic model, and shows why these strategies are effective though they do not invoke conventional NLP. The paper then argues that the statistical methods used for retrieval have potential applications in conjunction with NLP for other important tasks, taking automatic summarising as an example.

Keywords

Document / text retrieval, statistical methods, probabilistic model, natural language processing, summarising

1 Introduction: the task

Document, or text, retrieval is an important form of information retrieval. There are many variations in the task including, for example, recovering an imperfectly remembered but previously seen item; or seeking items for which some relatively unambiguous, non-content specification like an author's surname can be given. But the more important and more challenging case is where the user is seeking documents about some *topic* without prior knowledge of what there is in the file, or reliable pointers to it.

Adhoc searches of this kind may alternatively start from a given known document, seeking others like it. They may also subsume *iterative* searches in which both user need and search query are modified on the basis of what is retrieved each time. Iterative searching on some given occasion has much in common with topic-based information *filtering* and routing over a period of time: but I shall concentrate initially on adhoc searching and return to filtering later.

2 Retrieval essentials

Adhoc searching has two key features. First, however much a user already knows, they are by definition 'ignorant' or they would not be seeking more information. Second, whether or not the user's information *need* is expressed as a question, e.g.

`Is prefabricated furniture manufactured in developing countries?`

it should not be assumed that, even if they system was intelligent enough to answer the question, this would necessarily suffice. The actuality in many cases, and thus the default presumption, is that the user wants to find out about something and therefore wishes to be provided with materials on it for their own review and analysis. This is required to allow them to properly relate any new information to what they already know. The user's linguistic input is taken as a *request* for documents giving information on, i.e. about, the user's topic, for the user to read: e.g. documents on

`prefabricated furniture manufacture in developing countries.`

The frequently-made assumption is that since the user's request is a linguistic object - phrase, sentence, known document, i.e. a text or discourse of some sort, and each document in the file is also a linguistic object, natural language processing (NLP) is required for effective retrieval. Thus given that, as illustrated above, the input topic is a complex relational concept and, further, that this can be expressed in different ways, syntactic and semantic text analysis is needed both to determine the nature of the topic and to allow one expression of it to be related to another. We suppose, for instance, that the meaning of the example topic is found to be

'the making of household chairs etc that are in kit form when done in nations that are not economically advanced',

and that this allows the original expression of the topic to be seen as equivalent to, e.g.

`manufacturing prefabricated furniture in countries that are developing`

or

`developing country manufacture of furniture which is prefabricated.`

Adopted wholeheartedly, this view has an important consequence. It implies the ability to identify not just the presence of a topic in a document, but that a document is *about* a topic, if not exclusively at least to a substantial extent. Thus when the retrieval file contains full text documents as opposed to titles or abstracts, the implication is that document *index descriptions*, indications of what a document is about, are summaries. But making summarising a condition of indexing is a demanding strategy.

The fact that modern retrieval systems apply wholly different approaches to indexing and searching is not, however, because we cannot deliver good automatic summaries. These methods deliberately eschew syntactic and semantic processing of the conventional kind, and rely heavily on word-based statistics. They have been developed and tested, over a long period and on a large scale, in comparative experiments which have shown both that they are effective in their own right and that they perform as well as methods exploiting NLP (TREC, 1993-1998). From the NLP point of view these methods are astonishingly simple. But though it seems surprising that NLP-based indexing does not do better, the established statistical techniques are in fact well suited to the retrieval task. They are also more sophisticated than they at first appear. So their successful application in retrieval may have lessons for other language-based information processing tasks.

3 Statistically-based retrieval methods

Statistically-based retrieval strategies are underpinned by formal models, notably the vector space, probabilistic, and inference net models (Salton and McGill, 1983; Sparck Jones et al., 1998; Turtle and Croft, 1991). At the foundational level there are real differences between these. But as they all have to work from the same basic data and, for the good reasons detailed later, exploit these in much the same way, practical implementations, especially of the vector space and probabilistic approaches, look much the same. For present purposes it is not necessary to elaborate on the underlying theories. I shall treat one approach, the probabilistic one, as an example and, taking the theoretical account presented in Sparck Jones et al. (1998) for granted, use the following summary account of the form in which the model is practically applied (see e.g. Robertson and Sparck Jones, 1994/7; Okapi, 1997) to illustrate the key features of modern retrieval.

3.1 An illustration

We start with simple index terms, say single words. Requests and documents are represented, for indexing, just as sets of terms. In particular, a search *query* is a set of terms. These terms are *weighted* using three types of information, to determine the *matching score* between the query and a document. The first type of information is the frequency of a query term in a document, `TF`. In general, the more frequently the term occurs in a document, the more likely the document is to be about the query topic and hence *relevant* to the user's need. The second type of information is the document length, `DL`. Given two documents with the same length, if the query term is more frequent in the first than the second it is reasonable to conclude the first is more likely to be relevant than the second. But if the first document is longer than the second, the greater term frequency may be attributed to greater verbosity and not be especially useful. We therefor modulate term frequency by document length.

The third type of information is collection frequency, `CF`, the number of documents in which the query term occurs. Since retrieval is about selecting the (typically few) relevant documents from a mass of non-relevant ones, terms with high collection frequency are poor selectors, Further, their occurrence in a particular document may follow simply from this general high frequency. Collection frequency should thus be exploited inversely, and a query term's weight for a particular document reflect the extent to which `TF` and `CF` are interestingly related: a good query term, i.e. one to which a high weight is assigned, is one with high `TF` (subject to `DL`) and low `CF`, i.e. with high inverse `CF` as naturally represented on the logarithmic scale.

Taking these three types of information together thus produces a formula for the *combined weight*, `CW`, of a specific query term `i` occurring in a specific document `j` as follows:

```
CW (i,j)  =  [ CFW (i) * TF (i,j) * (K1+1) ] /

                        [ K1 * ( (1-b) + (b * (NDL (j)) ) ) + TF (i,j) ]
```

In this formula `CFW` defines a `CF`-based weight. Given `N`, the total number of documents in the file and (for later convenience) relabelling `CF` as `n`:

```
CFW (i)  =  log N - log n
```

`NDL` is a normalised document weight defined as:

```
NDL (j)  =  (DL (j)) / (Average DL for all documents)
```

K1 and b are tuning constants.

Query-document matching scores are therefore not based just on whether query terms are present in a document, but are the sum of the weights of the present terms. As discussed further below, there is no requirement to match on all the given query terms. But the use of weights allows more refinement than simply counting the number of matching terms, since a match on fewer, but good terms, can give a higher score than a match on more, but less good ones. The output of a search is thus a list of matching documents, *ranked* by score and, hopefully, with highly ranked documents relevant to the user's need.

The basic scheme just described can be modified in important ways. First, by using a fourth type of information: where requests are long (as they might be with example documents), the weighting formula can be adjusted to take account of the fact that terms may have different within-query frequency, since more frequent terms are presumably more valuable. This is done simply by multiplying the CW formula by QTF. The more significant modification is where there is some information about known relevant documents, as in iterative searching. Given this fifth type of information, individual terms can be characterised not just by their simple collection frequency, but by their respective relevance and non-relevance frequencies, and by the former, RF, in particular. Thus we first define a relevance weight RW for each query term i:

```
RW (i)  =  log [ ( (r+0.5)(N-n-R+r+0.5) ) / ( (n-r+0.5)(R-r+0.5) ) ]
```

where N and n are as above, R is the number of known relevant documents and r (i.e. RF) is the number of known relevant documents for the term, and 0.5 is included to allow for predictive uncertainty.

Then for iterative searching we can replace CW by an iterative weight CIW, defined as:

```
CIW (i,j)  =  [ RW (i) * TF (i,j) * (K1+1) ] /

                      [ K1 * ( (1-b) + (b * (NDL (j)) ) ) + TF (i,j) ]
```

Finally, it is reasonable to assume that other terms for which relevance information is available may be valuable additions to the query, given the variable ways that topics can be expressed. Query *expansion* increases the chance of matching. Thus if all the terms occurring in the known relevant documents are pooled, they can be ranked by offer weight, OW, i.e. by

```
OW (i) = r * RW (i)
```

and the top, e.g. 10 or 20, terms added to the request. The query terms are then weighted using CIW. This process is using a sixth type of information, which we can label co-relevance frequency, CRF. (It is of course also possible to combine QTF and CIW, in the obvious way.)

The effect on practice for the example topic might be that a term with low CF, typically high TF, and high RF, say prefabricated, will score very highly, perhaps ensuring that a document containing only prefabricated and furniture scores more highly than one containing furniture, manufacture, developing and countries.

It must be emphasised that the theory given in Sparck Jones et al. (1998) motivates this specific set of formulae: they are not just pulled out of the air as more or less plausible. Equally, it must be emphasised that applying them delivers a good standard of performance, compared both with other intuitively obvious techniques and with ones using NLP. Further, the other statistically-based models give very comparable performance, showing that it is right to use these types of information for retrieval purposes.

4 Topic and content

From the point of view of topic representation and (document) content capture, the essential feature of the modern approach to retrieval is most clearly seen in the simple unweighted case, and where there are more than one or two terms in the initial query, e.g.

`cheap prefabricated furniture manufacture`.

The key to modern retrieval is the *best match principle*. While, ideally, one would like to match all of the query terms, this is unlikely to be achieved in practice. Moreover, while any document matching on all terms where many are given is highly likely to be relevant, there will normally be other documents that are relevant although they do not match on all terms. Failure to get complete matches follows naturally from the uncertainties that characterise the whole situation:

1. partial expression of need (users don't know so much);
2. indirect access to content (words only say so much);
3. variable presentation of content (people don't coincide too much).

These uncertainties cannot be overcome by greater control aimed at fully explicit, normalising topic/content representation. This is most obvious in the first and second cases, but also applies to the third. Thus being about a topic is an essentially loose notion: there is no one correct way of being about something. Content representation and matching have therefore to be on a 'do as best you can' basis.

But while this tolerance, or hospitality, may seem attractive from one point of view, the obvious points against it are that matching on sets of terms does not guarantee that the correct term senses will be selected or, more significantly, that the correct structural relations between terms will be captured. We are all familiar with the claim that the strategy described is inadequate because it cannot distinguish a blind venetian from a venetian blind.

The reality is different. Just matching on several terms at once normally succeeds only for appropriate senses and the required structural relations, or at least for broad approximations to these. Conjoint word matching de facto reflects suitable sense selection conditions and semantic relationships. For example,

`cheap prefabricated furniture manufacture`,

though it may match documents that are about cheap furniture or cheap manufacture, is unlikely to encounter and hence match documents that are about the vulgar [i.e. cheap] contents [i.e.furniture] of someone's mind. This is clearly less likely when either some small subset of a large initial set of terms matches, e.g. `cheap` and `manufacture`, or when there are only a few terms to start with, e.g. `developing` and `countries`, which may match on quite independent occurrences of `countries` and `developing`. There are nevertheless successful matches even in these conditions, specifically ones which capture the intended concepts (and, we assume, therefore also retrieve relevant documents). The simple set model is indirectly capturing real complex concepts and finessing the need to specify these concepts explicitly, in full detail. Further, while matches may sometimes be inappropriate, retrieval is usually directed at obtaining sets of documents (say to cover a subject thoroughly), which means that while some matches may be duds, the user benefits from other successful ones. However the fact that in modern situations, e.g. World Wide Web searching, requests typically start with, maybe, only two terms suggests that, whatever other problems these short requests present, they offer little scope for NLP.

The more important point about partial matching is the real flexibility it offers: this is to permit retrieval even when the language or perspective of request and document are slightly different. Specifically, the set-based approach, especially where there are many query terms, as with expanded queries, assists with both precision and recall: the former by offering alternative forms of multi-term matching, the latter by achieving at least some match.

4.1 Key concept emphasis

The points just made apply to the basic unweighted case. But many tests have demonstrated that very large performance gains can be obtained with weighting, particularly when searching full text. Weighting clearly allows partial matching, just as the simple case does, but it does far more as well. The specific purpose of weighting is to meet the need for documents that are primarily, or substantially, about the query topic. Weighting signals that a term is *important*, because frequent, in a document. We do not know what it means, but this does not matter. Term weighting thus picks out the main concepts in the text and, where a high matching score is achieved, incidentally marks the fact that the query topic is important for the text. It does this somewhat crudely, just as set matching is a rather crude way of capturing structure. But it is sufficient for the purpose, even if it only generally, but not universally, holds. It also works in conjunction with partial matching.

4.2 Wider application

The successful approach to adhoc document retrieval just described can be extended to nearby related tasks. With filtering, for instance, as TREC has shown (TREC, 1993-1998), it is possible to apply the model very effectively when a large amount of relevance information is available. The same generic approach can be applied to passage retrieval, or short text extraction, and to the construction of new, hypertext documents (Salton et al., 1997). It has also been applied successully to other languages including character rather than word-based ones (e.g. Chinese in TREC), and to spoken documents, which may have distinctive discourse styles (as in broadcast news) as well as recognition errors.

The more interesting question is whether these strategies can be adapted to automatic document transformation and reduction, as in summarising, as opposed to document selection. I shall return to this after considering the role of NLP in retrieval in more detail.

5 NLP in retrieval

In spite of the points made in the last section, it is widely believed there is a manifest need to apply NLP for term and topic specification in retrieval, and attempts have been made to do this at least since the sixties (Sparck Jones, in press *a*).

The most basic, but also a very appropriate, use of NLP has been for word stemming, since there is good (though not universal) evidence that conflating e.g. singular and plural noun forms, or verb variants, is helpful in promoting recall without loss of precision. Stemming has been very effectively implemented with a procedure like Porter's (Porter, 1980), which uses a suffix list but no stem dictionary. It has been most widely exploited for English, but has also been applied to other languages, e.g. Spanish in TREC. In the original topic example, for instance, it could deliver the term `develop-` which would match `developing`, `developed`, `develops`, etc. Compounding languages require more effort for word decomposition and a reference lexicon, as well as more complex index management; but there is no reason in principle why the statistical methods described earlier should not be applied to the constituents of compounds.

This single term application of linguistic processing is nevertheless only minor, as well as uncontroversial. The belief in the value of NLP refers to a more ambitious application, to derive complex index descriptions. However the evidence from conventional manual indexing, where whole documents may be summarily indexed by complex proposition-like descriptions has shown that these are too demanding for effective retrieval, regardless of whether available NLP techniques can deliver them: see e.g. Keen (1973). Thus the potential use of NLP for retrieval has focussed on the identification and exploitation of compound terms, as represented by e.g. nominal groups. The argument here is that the components of topics may themselves be compound concepts, though it is sensible, in the interests of flexibility, to allow them to be used within the same set-based method of dealing with whole topics as is the norm for simple terms.

There are thus two issues: one is whether *compound terms*, often called *phrases*, are really effective; and the other is whether, if they are used, they have to be defined linguistically as opposed to statistically. The argument for the linguistic approach is illustrated by Strzalkowski (1994): parsing identifies head-modifier units, which may then be conflated, for the same reason as with word variants, and also stemmed, so that e.g. the original `furniture manufacture` will also match `manufacture of furniture`, `furniture manufacturing` and `furniture that is manufactured`. Unfortunately, while there is some support for the claim that phrasal terms contribute usefully, if not spectacularly, to retrieval performance, there is no evidence that there is any gain from using NLP as opposed to statistical techniques (Mitra et al., 1997). In fact, just as with the set combination of single terms, a pair of words that recurs in a file is likely to represent a genuine compound, which can be identified by this recurrence without the need for parsing. This strategy could, for instance, pick up the adjacent pair `prefabricated furniture` along with, after stemming, `prefabricating furniture`, and also `furniture prefabrication`. (In line with the general need for flexibility, the elements of phrasal terms are used in requests as well as the phrases: the latter tend to have higher weights, so are more valuable, but the former contribute to matching too).

6 Lessons for NLP

The main lessons to be learnt from experience with the document retrieval task are first, the fact that at least some language-using tasks can be effectively done in a very crude way. This is not only because they are actually rather undemanding tasks. The second and more important lesson is that these tasks can be effectively done in a crude way because language, i.e. discourse, is *redundant*. The whole approach depends on repetition and reinforcement, with the same concepts (words) figuring in multiple slightly different ways in a text. It is no accident that frequency signals importance. NLP workers have always recognised that corpus data is valuable e.g. for grammar refinement or lexicon tailoring, precisely because it captures significant differences, marked by relative frequency of coccurrence and cooccurrence, in usage. However the point that emerges from the experience of statistically-based retrieval is that is that it may be worth trying to extend the application of statistical methods further in building tasks systems. In particular, retrieval suggests that statistical data should be exploited not just to improve the knowledge resources that are applied, but in a stronger sense, motivated by some underlying model such as the probabilistic one, to drive task processes.

Automatic summarising is a critical test for this. Attempts in the past to derive summaries by extracting and concatenating source text sentences, chosen using word frequencies, have not been convincing (Paice, 1990). But as summarising comes in many flavours for many purposes (Sparck Jones, in press *b*), there may be a role, for instance, for 'semi-summaries' consisting of statistically-grounded phrases. This would not just be a re-run of retrieval methods, first because phrases would be chosen to be representative of the individual document alone, not to discriminate one document from others. More importantly, it would be necessary to combine NLP of the conventional kind with the use of frequency data: first because when confined to a single document it is necessary to analyse it more carefully than for retrieval, since it is not possible to rely so heavily on file information to identify important phrases; second to capture genuine phrasal variants as fully as possible; and third because for individual documents it is desirable to find long phrases, since these are more informative for the reader and there is no need to consider the matching constraints that make shorter phrases more helpful in the retrieval case.

The kinds of possibilities that are beginning to be explored are illustrated in e.g. Mani and Maybury (1997, in press). Some Cambridge work also shows how the general suggestion that NLP and statistical methods might be helpfully combined can be applied at different levels of analysis: more simply at a very shallow level, or at a somewhat deeper level, with less reliance on the form of source expressions.

In a modest study along the first lines, Dersy (1996) used robust parsing to identify e.g. noun groups, conflation of variants, phrase selection based on occurrence frequency but constrained by rules designed to avoid phrases with heavy lexical overlap, and generation of an output list of phrases in preferred surface form and in text order. Tucker's recent work (see Sparck Jones, in press *b*), on the other hand, seeks to escape surface linguistic expressions and make more explicit use of the discourse relationships between source text concepts. Thus in his approach, sentences are parsed into logical forms from which primitive predications are extracted. These are connected e.g. by predicate or argument sharing and thus form a cohesion graph embodying the overall attentional structure of the source text. Operations on the graph are designed to select subsets of nodes taken to represent the main source content, applying criteria e.g. for salience, representativeness and coverage of the original. Predications in the selected subset are merged where possible, by using e.g. common arguments, and and an output list of surface phrases is then generated to provide, as with Dersy, an indicative summary. In contrast to Dersy, however, the merging can lead to complex and thus more informative, phrases. Evaluation work, though difficult and hence limited, suggests this type of approach to summarising can be of considerable value for browsing and preview purposes, though such summaries cannot take the place of 'proper' condensation and reformulation, based on source text understanding and delivering a coherent output text.

Similar methods would also appear to be applicable to information extraction, as general strategs designed to identify key source text elements without relying on prior specification of what is required (e.g. terrorist incidents, company merger information).

Current NLP techniques for robust parsing are sufficiently effective to act as one of the necessary supports for identifying important types of text unit. My claim here is that they may have a natural partner in current retrieval techniques. These have become well-established in their own area: the nature of their success there suggests they may be more widely relevant to other language-processing tasks than hitherto supposed, so further research on combining them with NLP is well warranted.

References

Dersy, J. (1996) Unpublished MPhil Dissertation, University of Cambridge.

Harman, D.K. (1996) Evaluation techniques and measures. *The Fourth Text REtrieval Conference (TREC-4)*, (Ed. D.K. Harman), Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, 1996, A6-A14.

Keen, E.M. (1973) The Aberystwyth index languages test. *Journal of Documentation*, 29 (1), 1-35.

Mani, I. and Maybury, M. (Eds. (1997) *Intelligent scalable text summarisation*, Proceedings of a Workshop, Somerset NJ: Association for Computational Linguistics.

Mani, I. and Maybury, M. (Eds.) *Advances in automatic text summarisation*, Cambridge MA: MIT Press, in press.

Mitra, M., Buckley, C., Singhal, A. and Cardie, C. (1997) An analysis of statistical and syntactic phrases. *Proceedings of RIAO-97, Computer-Assisted Information Searching on Internet*, Centre de Hautes Etudes Internationales d'Informatique Documentaires, Paris.

Okapi (1997) Papers on the Okapi system, Special Issue of *Journal of Documentation*, 33, 3-87.

Paice, C.J. (1990) Constructing literature abstracts by computer: techniques and prospects, *Information Processing and Management*, 26 (2), 171-186.

Porter, M.F. (1980) An algorithm for suffix stripping. *Program*, 14, 130-137.

Robertson, S.E. and Sparck Jones, K. (1994/7) Simple, proven approaches to text retrieval. Technical Report 356, Computer Laboratory, University of Cambridge.
(http://www.cl.cam.ac.uk/ftp/papers/reports/TR356-ksj-approaches-to-text-retrieval.ps.gz)

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

Salton, G., Singhal, A., Mitra, M. and Buckley, C.. (1997) Automatic text structuring and summarisation. *Information Processing and Management* 33 (2), 193-207.

Sparck Jones, K. (in press *a*) What is the role of NLP in text retrieval? In *Natural language information retrieval* (Ed. T. Strzalkowski), Dordrecht: Kluwer.

Sparck Jones, K. (in press *b*) Automatic summarising: factors and directions. In *Advances in automatic text summarisation* (Ed. I. Mani and M. Maybury), Cambridge MA: MIT Press. (http://xxx.lanl.gov/cmp-lg/9805011)

Sparck Jones, K. Walker, S. and Robertson, S.E. (1998) *A probabilistic model of information retrieval: development and status.* Technical Report 446, Computer Laboratory, University of Cambridge.
(http://www.cl.cam.ac.uk/ftp/papers/reports/TR446-ksj-probabilistic-information-retrieval .ps.gz)

Strzalkowski, T. (1994) Robust text processing in automated information retrieval. *Proceedings of the 4th Conference on Applied Natural Language Processing*, Somerset NJ: Association for Computational Lingustics, 168-173.

TREC (1993-1998): D.K. Harman (Ed.) *The First Text REtrieval Conference (TREC-1)*, Special Publication 500-207, National Institute of Standards and Technology, Gaithersburg, MD, 1993; *... Second ... (TREC-2).* SP 500-215, NIST, 1994; *... Third ...(TREC-3)*, SP 500-225, 1995; *... Fourth ... (TREC-4)*, SP 500-236, 1996; Voorhees, E.M. and Harman, D.K. (Eds.) *... Fifth ... (TREC-5)*, SP 500-238, 1997; *... Sixth ... (TREC-6)* (1997), SP 500-240, 1998.

Turtle, H. and Croft, W.B. (1991) Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9 (3), 187-222.