DOCUMENTATION NOTE

QUERY-DOCUMENT SYMMETRY AND DUAL MODELS

S.E. ROBERTSON

Centre for Interactive Systems Research
Department of Information Science, City University
London EC1V 0HB

The idea that there is some natural symmetry between queries and documents is explored. If symmetry can be assumed, then it leads to a conception of 'dual' models in information retrieval (given a model, we can construct a dual model in which the roles of documents and queries are reversed). But symmetry breaks down in various ways, which may invalidate this construction. If we can construct a dual, it is not obvious that it can be combined with the original.

IN THEORETICAL AND EXPERIMENTAL PAPERS on information retrieval, the basic concepts concerned (documents/records, information needs/requests/queries, words/index terms etc.) and the various relationships between them are treated in a more or less formal way, and assumed to have certain formal properties. One particular property which appears in some cases is one of symmetry, as between documents and queries. (An even stronger property sometimes attributed, which implies symmetry, is identity: the assumption that documents and queries are actually the same kind of entity.) Not all approaches assume or imply symmetry, let alone identity; however, symmetry or identity is often implicit in particular approaches. Some examples are discussed below.

In this note, I propose to explore the notion of symmetry and some of its consequences, and discuss how well it reflects reality.

1. SYMMETRY AND DUALITY

In an information retrieval system, documents and queries are represented by means of some language. The relevance relation is a relation between documents and queries, and the function of the system is to predict this relation by means of some matching function which matches the representations of documents and queries.

Journal of Documentation, vol. 50, no. 3, September 1994, pp. 233–238

This definition is perfectly symmetrical, in that we could change every occurrence of the words 'document' and 'query' to 'query' and 'document' respectively without in any way changing the sense. Thus if the definition is accepted as a complete (for formal purposes) specification of what an IR system is, then we must also accept (again, for formal purposes) that the situation *is* symmetrical. We may illustrate the idea with reference to two different views of information retrieval, one apparently symmetrical and the other not.

The vector space model [1], at least in its simplest form, regards both documents and queries as weighted sets of index terms, which may be represented as points in an n-dimensional vector space, where n is the total number of available index terms. (This view is not only symmetrical, it is an example of an approach which sees documents and queries as essentially identical objects.) On the other hand, an approach which starts from the idea that queries are Boolean (and documents are sets of index terms or continuous text) appears to be asymmetric.

It should be pointed out that the interpretation of an approach as being symmetrical or otherwise is not necessarily straightforward. For example, within the vector-space approach, concepts may be introduced which appear not to be treatable symmetrically: an instance would be passage retrieval, where the document is assumed to have a partitioned or possibly hierarchical structure of sections. On the other hand, in the second example there is no *necessary* reason why one should dismiss the possibility of representing *documents* by Boolean expressions.

If, then, we have an approach to IR or IR systems which is essentially symmetrical, and a model or theory which may be asymmetrical, we can construct a dual model in which the roles of the documents and queries are reversed. There follow two brief examples of such duality.

1.1 Probabilistic indexing and searching

In 1960, Maron and Kuhns [2] put forward a theory of probabilistic indexing; in 1976, Robertson and Sparck Jones [3] proposed a model of probabilistic searching. It subsequently emerged [4] that the two models are in fact dual, at least at some level of abstraction, in the following sense.

The indexing model assumes that search terms are given (that is, the searcher comes with an information need which she or he expresses in a certain way, irrespective of the system); it is then the function of the system (including the indexers) to ensure that documents are indexed in appropriate ways, so that the system responds to the request in a good way. The searching model, on the other hand, assumes that the indexing terms or document representations are given; the system's function is to ensure that a suitable search formulation is made. This situation is not obviously symmetrical; however, it becomes so if it is assumed that both documents and queries arrive with some fixed representation, but in both cases the system may make its own extra associations in order to bring the two together. The two models are then clearly dual to each other (Robertson, Maron and Cooper [4]). (As with the vector-space approach

discussed above, developments of either model may introduce aspects which are or appear asymmetrical.)

The problem considered in Robertson, Maron and Cooper [4], of how to combine the two models, is discussed further below.

1.2 2-Poisson model

The second example is discussed by Robertson *et al.* [5], and involves the application of the 2-Poisson model of term frequencies. This model has been used by a number of authors in information retrieval, in connection with within-document frequencies: it assumes that the distribution of within-document frequencies of a given term over the document collection is a mixture of two Poisson distributions. In Robertson *et al.* [5], it is applied first to the documents in the usual way, and then to the queries; that is, it is assumed that the distribution of within-query term frequencies (in textual queries) for a given term over the query collection is a mixture of two Poisson distributions. This latter model is the dual of the former.

Again, there is a problem regarding the combination of the two models, discussed further below.

2. OBJECTIONS TO SYMMETRY

Thus if we can assume the property of query-document symmetry, we can make use of duality to generate new models – potentially a powerful device. However, it is worth enquiring as to the validity of the symmetry property, by considering objections to it. There are in fact several points which cast doubt on the idea.

(a) One way of seeing the function of the system is to rank the documents for presentation to the user, without making an explicit, separate prediction of relevance for each one.

This ranking is definitely asymmetric: the system is *not* required to rank the queries in relation to each document. It could be argued that the specific prediction is achieved by the choice of a cut-off point on the ranking, but if this is determined by the user rather than by the system, symmetry is not restored.

It may also be argued that the traditional, retrospective retrieval task (stable database of documents, new query to be processed) looks like this, but the routing (or filtering or SDI) task (stable database of queries, new document to be processed) provides the symmetrical counterpart. If this is so, then a model applicable to retrospective searching would have a dual applicable to routing (and vice versa). However, it is not the case in routing that the system can simply rank the queries in relation to the incoming document – in this case, the system does have to make an explicit prediction about each query.

We may then argue that objection (a) above does not apply to the routing task alone, although it may apply to retrospective searching.

(b) Relevance judgements are made by the requester, not by anyone representing the interests of the document (e.g. the author).

It could be argued that relevance judgements can (and perhaps should) be made by independent judges (though this is normally regarded as second best). But in any case, the judge will normally be trying to assess what is good *for the question* (or for the information need or the problem that gave rise to the request), and not for the document. There may, however, be some exceptions: in the routing case, for example, it is conceivable that a relevance judge's task would be to determine who is the best person to deal with the document (on the assumption that there is a management requirement that every document should be dealt with). Again, if the task is to distribute a set of papers submitted to a conference among a set of referees, this objection does not apply. But it is clear that not all situations are symmetrical in this respect.

(c) A document is a document is a document; a request is merely a (partial) representation of an information need.
(d) Relevance judgements match actual, explicit documents with implicit information needs.

Again, one can imagine situations to which these objections do not apply; however, they do apply in many (if not most) IR situations.

(e) Even if logical symmetry is accepted, we cannot assume statistical symmetry – it is very clear that, in general, the statistical characteristics of queries are different from those of documents.

This objection applies to any assumptions one might want to include in a model concerning, for example, statistical independence or specific statistical distributions. This is certainly an open question in regard to the 2-Poisson models discussed above – that is, even if the 2-Poisson assumption is applicable to the documents, and the logic of the situation allows us to generate the dual model, there is no implication that the statistical assumption will also work in the dual situation. (As it happens, it appears to work rather better in the dual model!)

The question then arises: do these various objections invalidate any argument based on symmetry? The answer is that it must depend on the model. Some models address only properties of queries and documents for which symmetry holds (any model is an abstraction, and the particular abstraction may simply ignore – not be concerned with – those properties to which the objections apply). Within such models, the duality argument looks useful. But one must certainly be wary of applying it indiscriminately.

3. COMBINING A MODEL WITH ITS DUAL

Even if we accept a symmetry argument in a particular case, and use it to generate a dual to any particular model (by simply reversing the roles of documents and queries), it is not obvious that the pair of models can properly be combined. This problem will be illustrated with the two examples given earlier.

3.1 Probabilistic indexing and searching

The specific combination problem for probabilistic indexing and searching was discussed extensively by Robertson, Maron and Cooper [4]. Essentially, one model assumes that the usage of terms in queries is given and determines their use in indexing documents; the other assumes the reverse. A combined model should determine their use in both cases, but would then find that nothing is given: terms would be unattached to anything, so nothing could be inferred about them.

The approach proposed in [4] clarified the symmetry issue and allowed a resolution of this conflict. It was, as indicated in section 1.1 above, to separate document properties (which might be terms) from query properties (which might also be terms). Both kinds of properties are taken as given, but the model is allowed to infer simultaneously the relation of queries to document properties and of documents to query properties.

3.2 2-Poisson model

There appears to be a similar problem in the case of the 2-Poisson model.

The within-document term frequency model can be described as follows (Robertson *et al.* [5]). We take a given query term, and assume first that its frequency of occurrence in documents reflects a hidden variable. This is a binary variable known as eliteness (to the term in question); documents are either elite or not.* Secondly, we assume that this document property of eliteness to the term is associated, in a statistical sense, with relevance to the query. This second assumption makes sense only because we know the term to be a query term.

The dual model, then, is as follows. We take a given document term, and assume (a) that its frequency of occurrence in queries reflects a similar hidden variable (which we will continue to call eliteness), and (b) that this *query* property of eliteness is associated with relevance to the document. (Although 'relevance of a query to a document' is a somewhat unusual way to refer to the relevance relation, it is just that – another way of describing relevance in the usual sense – at least if we can assume symmetry.) Again, the second assumption only makes sense because we know the term to be a document term.

In order to combine the two models, however, we must abandon such certainties. If the frequencies of occurrence (of any term we care to consider) in queries and in documents reflect, in both cases, hidden variables, then it is by no means so obvious what kind of relation we may be able to assume between these hidden variables and relevance of the document-query pair.

This question has not yet been resolved.

* The property of eliteness might be interpreted as being *about* the concept represented by the word. It is term-specific (in a multi-term query there will be an eliteness property associated with each term separately), but it is not query-specific. The matter clearly deserves further discussion than is possible in this paper; Robertson and Walker [6] take it a little further.

4. CONCLUSION

The assumption of symmetry between queries and documents leads to a potentially powerful concept of dual models. However, there are two problems with this concept. Firstly, symmetry breaks down in various ways; it can therefore be assumed only when the ways in which it breaks down are not relevant to the models under consideration. Secondly, how to combine a dual pair of models is not at all obvious.

ACKNOWLEDGEMENTS

I am very grateful to two referees for comments on an earlier version of the paper.

REFERENCES

1. RAGHAVAN, V.V. *and* WONG, S.K.M. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science, 37*, 1986, 279–287.
2. MARON, M.E. *and* KUHNS, J.L. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7, 1960, 216–244.
3. ROBERTSON, S.E. *and* SPARCK JONES, K. Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*, 1976, 129–146.
4. ROBERTSON, S.E., MARON, M.E. *and* COOPER, W.S. Probability of relevance: a unification of two competing models for information retrieval. *Information Technology – Research and Development, 1*, 1982, 1–21.
5. ROBERTSON, S.E. *et al.* Okapi at TREC–2. *In*: HARMAN, D.K., *ed. The Second Text REtrieval Conference (TREC–2), Gaithersburg, August 1993.* Gaithersburg, MD: NIST, 1994, 21–34.
6. ROBERTSON, S.E. *and* WALKER, S. Some simple effective approximations to the 2-Poisson model for probabilistic information retrieval. *In*: *SIGIR 94: Proceedings of the 17th International Conference on Research and Development in Information Retrieval, Dublin, July 1994.* Forthcoming.

(*Revised version received 17 March 1994*)