Searching via Keywords or Concept Hierarchies – Which is Better?

Richard F. E. Sutcliffe[1], Kieran White

Department of Computer Science and Information Systems,
University of Limerick, Limerick, Ireland
{Richard.Sutcliffe,Kieran.White}@ul.ie

Abstract

We have carried out a comparison of interactive search in a homogenous information retrieval domain using a keyword search engine on the one hand and a concept ontology system on the other. The experimental design was that of the TREC Interactive Track. While the results showed that keyword search was superior on this occasion, we have identified the ideal characteristics of an ontology and shown that the one used for the study did not conform to these. Future work will include repeating the experiment with an optimal hierarchy and establishing numerical attributes of an ontology relative to a particular task domain.

1. Introduction

Two approaches to information retrieval are to search via keywords and to traverse a hierarchy of concepts which gradually become more specific until documents relating to the user's information need are found. In the former case, a user transforms their information need into a short query which is entered into the system. An ordered list of matching documents is then returned. The user then inspects the content of each looking for the answer to their query. In the latter case, a user compares their information need to a set of topic descriptors, usually specified as short text strings. They then click on the one which matches best, leading to the presentation of a new set of descriptors. The process is repeated until a set of documents is reached. These must then be inspected in turn.

The objective of this work was to make a direct comparison between these two methods when applied to the task of retrieval in a homogenous task domain – word processing. Additional objectives were to investigate users' attitudes to different search methods, to establish the characteristics of an ideal taxonomy and to experiment with interactive methods of retrieval evaluation.

In the next section we outline previous work on ontology construction and evaluation before turning to the details of our own study.

2. Previous Work

2.1 Construction of Ontologies

Methods for the construction of ontological search engines are not the prime focus of this paper, which is concerned with the evaluation of such systems. However we make some brief remarks about them here. There are essentially three approaches which can be taken. Firstly, documents can automatically be organised into a hierarchy which can then be searched by a user. van Rijsbergen (1979) gives a thorough survey of such methods which indicates that they are not new to information retrieval. Many of these methods use a means of measuring the distance between two concepts together with an approach to clustering using that distance. Much subsequent work has been based on these ideas. For example the clustering method of the Scatter/Gather system uses a conventional document distance measure along with a partitional clustering scheme which can produce a hierarchy while being more efficient than binary agglomerative schemes (Pirolli et al., 1996; Hearst and Pedersen, 1996). The approach of Sanderson and Croft (1999) adopts a different hypothesis, namely that an ontological link between terms can be posited if a subsumption relation exists between them: x is the parent of y if the documents in which y occurs are a subset of those in which x occurs.

In considering automatic techniques, it should be borne in mind that the automatic construction of hypertext documents is related and has been the focus of a number of studies (see Smeaton and Morrissey (1995) for a review).

A second approach is to create a concept hierarchy by hand and then to develop a method for linking documents to it automatically. We discuss work we have undertaken with one such system in this paper. Finally, it is possible to attach documents by hand to a manually created ontology. This is a method commonly used for creating directories on the World-Wide Web.

2.2 Evaluation of Ontologies

Pirolli et al. (1996) undertook a study which aimed to determine judgements of clusters produced by their Scatter/Gather system. 16 experimental subjects were used together with 12 topics (queries) taken from the TIPSTER collection used in TREC. The main purpose was to make a direct comparison between the performance of Scatter/Gather and a keyword search engine. Subjects were also asked to estimate the percentage of texts relevant in a Scatter/Gather cluster, to create query terms based on an analysis of clusters, and to draw hierarchical diagrams based on Scatter/Gather output. The experimental design was similar to that used in the TREC Interactive Track (Hersh and Over, 2001). The main result of the study was that subjects who used Scatter/Gather alone to answer queries were twice as slow as those using a search engine. However the study also showed that the system was helpful in allowing users to formulate good queries and to understand the structure of the text collection.

Sanderson and Croft (1999) evaluated their system by asking a set of eight subjects to judge parent-child

[1] For related papers etc see www.csis.ul.ie/staff/richard.sutcliffe

relationships taken from a set of ontologies, each one automatically-generated from a particular TREC topic. Their intention was to find out how valid the ontologies were, rather than how efficiently retrieval could be carried out with them. The main finding was that 48% of automatically created parent-child pairs were judged 'interesting' as compared with 28% of randomly generated pairs.

3. Systems to be Compared

The first system in the study (System A) is a conventional keyword-based retrieval engine using inverted indexing and the vector space model. Searching involves typing in a short query, receiving in response an ordered list of document identifiers and then inspecting the contents of promising documents.

The second system (System B) uses a manually-created concept hierarchy. Each node in the hierarchy has associated with it a set of conditions (in terms of keywords) which a document must satisfy if it is to be attached to that point in the hierarchy. Searching involves inspecting the top-level list of categories, selecting the category most appropriate to the information need, choosing the most appropriate category underneath this, and so on until the category closest to the topic of the query is found. Promising documents attached to this category can then be inspected. System B also supports search of the taxonomy itself (not the documents attached to it) via keywords.

Both System A and System B are commercial packages which were developed elsewhere. System B requires a concept hierarchy which is tailored to the application domain in order to achieve optimal results. This was created for us especially for the study.

4. Method

The experimental design used for the study was a version of that adopted in the TREC Interactive Track (Hersh and Over, 2001):

16 respondants are divided into two groups of eight, Group A and Group B. There are sixteen respondant packs. Each one is different. The pack consists of tutorial material on each system, questionnaires and the questions to be answered by a particular respondant.

Groups A and B participate in different sessions. Members of Group A are given an explanation of the task, a tutorial on System B, four questions on System B, and a questionnaire on System B. This is followed by a tutorial, four questions and a questionnaire on System A. At the end there is an exit questionnaire. The session for Group B is the same except they work first with System A and then System B.

Each of the two sessions lasted 88 minutes and was conducted according to the timetable of Figure 1.

The core of the task is the four questions on each system. Respondants are given six minutes per query to find any documents which are relevant to it. For each such document found they write down its number and the current clock time.

The organisation of queries, respondants and systems is intended to control for the effect of differences between respondants, the varying difficulty of queries, the effect of experience with one system on the use of the other, the effect of experience with a given system on its use with a later query, and any interaction between a particular query and a particular respondant.

Activity	Minutes
Welcome, Task Explanation	5
Tutorial on System B	10
Four Queries on System B	24
Questionnaire on System B	5
Tutorial on System A	10
Four Queries on System A	24
Questionnaire on System A	5
Exit Questionnaire	5
Total	**88**

Table 1: Timetable for the evaluation experiment. Session One is shown. For Session Two, A & B are exchanged.

5. Materials

The application domain for the study was word processing using the Lotus Ami Pro program. This was chosen as we had previously collected a set of 572 queries for Ami Pro and determined the answers relative to sections in its instruction manual in both English and Japanese (Sutcliffe and Kurohashi, 2000).

For this study eight queries were selected from the 572. These are shown in Table 2.

Both systems indexed the manual assuming that each section in each chapter was to be considered a separate 'document'. The results of a query in System A therefore comprised an ordered list of hyperlinks to manual sections. Similarly in System B any node in the taxonomy had attached to it a list of hyperlinks to manual sections each of which met the keyword-based conditions for being associated with that node.

1	How does Ami Pro sort characters which are neither alphabetic nor numeric?
2	Insert an equation in a table cell?
3	What is "Paste Special"?
4	How do I use various bullet types in a paragraph style?
5	If I wish to graph data how do I do this?
6	Printer setup?
7	Does the grammar checker correct misspellings?
8	Is it possible to number pages automatically?

Table 2: The eight Ami Pro queries used in the study.

6. Results

A correct answer to a query was taken to be a section which was considered relevant to that query by three or more out of five respondants in an earlier elicitation process which followed and refined upon Sutcliffe and Kurohashi (2000).

The Recall of a particular respondant working on a particular question was defined to be the number of correct answers written down divided by the total number of correct answers. Precision was not computed and hence incorrect answers written down were not taken into

consideration.

Using the individual Recall values for respondant-query-system combinations, the Mean Recall for each query-system combination was computed and hence the Average Recall for each system.

Query	System A	System B
1	0.50	0.25
2	0.38	0.29
3	0.88	0.79
4	0.47	0.47
5	0.34	0.06
6	1.00	0.88
7	0.38	0.40
8	0.50	0.31
Total	**0.55**	**0.43**

Table 3: Results in terms of Mean Recall per query and system.

The results are shown in Table 3. As can be seen the overall performance of System A was 0.55 and that of System B was 0.43. In other words, the system based on keywords performed better than the one using a taxonomy. On looking at the queries individually it can be seen that for Query 7, System B had a superior Recall. Using the times written down by respondants, the average time to find a correct section for this query was computed for both systems. The results were System A: 2.53 minutes and System B: 1.88 minutes. In the case of this query, therefore, the taxonomy based system was considerably faster.

7. Discussion

7.1 Overall Results

The first point to note about the results is that both systems performed poorly. Even though each query was short and clear, only just over half the correct answers were found on average, even by the better system. This suggests the application domain is a hard one. There are two reasons for this. Firstly, it is homogenous, meaning that all documents and queries are, in information retrieval terms, very similar. The existence of a term is not sufficient to indicate the topic of a passage, only its precise context of use. Secondly, word processing or indeed any work involving a computer package is by its nature exacting. Moreover, users require a correct and complete solution. In Information Retrieval terms this equates to the need for very high Precision and Recall together with low search times. This situation can be compared to a heterogenous domain in which documents deal with many topics. In such a domain, users may be satisfied with a system even though Precision and Recall are low, simply because *some* information in relation to a topic can always be found, whatever the topic.

The second point regarding the results is that in general, for our experimental setup, keyword search was found to be superior to taxonomy-based search. However, our experiences with taxonomy systems suggest that the performance they give is heavily dependent on the quality of the taxonomy and its appropriateness to the domain. We identified the following characteristics as being important:

- It must be possible to decide easily which category at a level is relevant to the query;
- Categories should ideally be mutually exclusive;
- The level of branching at any level must be limited;
- The depth of the ontology must be restricted;
- The number of documents attached to leaf nodes must be small.

We will now discuss these characteristics in turn. When an ontology is being used, the query concept must be mentally compared with all the possible successors to the current node in the ontology and an appropriate one chosen. This is a difficult task for the user and must be carried out on the basis of a short textual description which therefore needs to be clear. If it is not possible to make a decision rapidly and easily, the advantage of an ontology search is lost. The second point relates to the mutual exclusivity of choices. If several branches partially apply at a particular point in the search, then each must be tried in turn. As noted earlier, this is a very difficult task for humans who are not adept at taxonomy traversal. Therefore, exactly one choice should apply.

Turning to the level of branching, the more successors which exist at a particular stage in the search, the more time and effort is involved in deciding between them. Above a certain point, the user will become annoyed with the system and revert to a keyword search. The same point applies to the depth of the ontology. As more and more levels of the ontology are traversed, the user will become increasingly impatient for an answer.

The final point concerns the number of documents at leaf nodes. Ideally there should be just one document (or one answer) at each leaf. If there is a long list of documents, the taxonomy traversal becomes more like an exhaustive search of the document collection.

Unfortunately, the ontology used in our study did not satisfy the criteria discussed above. In particular, respondants reported two problems:

- Difficulty in deciding on an appropriate category;
- Finding an excessive number of documents at leaf nodes, leading to an exhaustive search.

7.2 Style of Usage of Systems

In considering the way in which the systems were used in the task we concluded that there are two broad patterns. If a system is used non-deterministically, the user makes an attempt to find a solution; if this does not succeed, they backtrack to a point in the search process at which a choice was made and select another alternative. In the context of a keyword search engine, different documents returned by a system can be inspected, based for example on the degree of match or on their outline descriptions. If this strategy does not succeed, the user can return to the query and re-formulate it. In the context of an ontology search, a previous decision regarding the choice of successor at a particular point in the hierarchy can be re-visited and amended.

An important finding of this study is that keyword searches are amenable to non-deterministic search while ontology searches are not. The reason for this might be that keyword searches are intrinsically short and bushy –

an amendment to the query is made and the results rapidly assessed before the nature of the amendment is forgotten. By contrast, ontology searches can be much deeper, in a tree which may still be bushy. While it is easy for a search algorithm to traverse a tree systematically, people are very poor at this task.

To put this finding in a different way, if an ontology search does not lead directly to a correct answer it is essentially a complete failure. By contrast, if a keyword search does not immediately succeed, the query can be modified without excessive effort on the part of the user.

7.3 Comments on the Evaluation Approach

The experimental architecture used in this study enjoys a number of advantages: Firstly, it is well established through its use in TREC Interactive Track. Secondly, it controls for a number of effects which we wish to eliminate, such as differences between the level of experience of respondants and variations in the difficulty of queries. Thirdly, it provides within its terms of reference a close estimate of the differences between two systems. Finally, it only requires respondants to be present at one session which is convenient from an organisational perspective.

On the other hand, it also suffers from some disadvantages: Firstly it is rather unwieldy – the overall time spent by subjects in the session is quite large and at the same time the amount of time devoted to each query is relatively small. With modest increases in the number of queries or the time per query, the overall session can quickly become excessively long resulting in loss of concentration by respondants. For the same reason, there is only a limited amount of time available for training. In our case only ten minutes could be allocated to each system tutorial. Secondly, two respondants are required for every query. This means that very few queries can be used in a particular study. These may not therefore be representative of the problem domain. While only eight queries could be used in the interactive study reported here, the full set of 572 can be exploited in an off-line evaluation. Thirdly, a large of group of respondants must be recruited who are not only willing to undertake the study but also have sufficient background knowledge of the subject matter itself as well as a basic understanding of computer-based search engines. In the case of word processing, undergraduate students satisfied these criteria quite well. However, this might not be so if the domain was more specialised.

Fourthly, the type of information which can be gathered regarding the processes underlying a person's search strategies is fairly basic, being essentially limited to candidate answer sections. Finally, the experimental design does not allow an absolute measure of a system's worth be computed. In consequence, comparisons must always be between pairs of systems.

8. Conclusion

Our study was a useful first step in learning about the characteristics of ontological search systems and how to evaluate them. For the time being, our results are inconclusive regarding the benefits of such systems. We are therefore planning to undertake the following next steps:

Firstly, we wish to re-run the study with a System B taxonomy satisfying our criteria to the maximum extent possible. In particular this would imply a low level of branching and minimal numbers of documents at leaf nodes.

Secondly, we would like to compare the results with the use of a specially tailored taxonomy based not on the documents but on the queries. This would enable the optimum performance of an ontology system to be established for this domain.

Thirdly, we should try the study on a different domain, for example one which is heterogenous. In general it is desirable to establish the kinds of application in which taxonomy searches can work effectively, in other words the ones on which the criteria discussed above can all be satisfied. In principle, homogenous domains should be good candidates because exhaustive listing of user problems may be possible.

Finally, the ideal characteristics for a taxonomy were only qualitatively phrased (e.g. depth 'limited' etc). We need to establish exact figures so that metrics can be developed for determining the efficacy of an ontology before user evaluation takes place.

9. References

Hearst, M.A., and Pedersen, J. O., 1996. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. *Proceedings of the 19th ACM SIGIR Conference, Zurich, August 1996*.

Hersh, W., and Over, P., 2001. TREC-9 Interactive Track Report. In D.K. Harman (Ed.), *The Ninth Text REtrieval Conference (TREC-9)* (pp. 41-50). Gaithersberg, Maryland: National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-249. Available electronically at trec.nist.gov.

Pirolli, P., Schank, P., Hearst, M.A., and Diehl, C., 1996. Scatter/Gather Browsing Communicates the Topic Structure of a Very large Text Collection. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), May 1996.*

Sanderson, M., and Croft, B., 1999. Deriving Concept Hierarchies from Text. *Proceedings of the 22nd ACM SIGIR Conference, Berkley, CA*, 206-213.

Smeaton, A.F., and Morrissey, P.J., 1995. Experiments on the Automatic Construction of Hypertext from Text, *The New Review of Hypermedia and Multimedia: Applications and Research, Vol 1, 1995.* http://lorca.compapp.dcu.ie/~asmeaton/pubs-list.html

Sutcliffe, R.F.E., and Kurohashi, S., 2000. A Parallel English-Japanese Query Collection for the Evaluation of On-Line Help Systems. Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May - 2 June, 2000, 1665-1670.

van Rijsbergen, C.J., 1979. *Information Retrieval* (second edition), Chapter 3. London, UK: Butterworths.

Acknowledgements

Many thanks to Bill Hersh of Oregon Health Sciences University for advice relating to the design of the study, and to Alan Buckeridge, Ruth O'Donovan and Darina Slattery for help with preparation of materials and for commenting on an earlier version of the paper.