*Information Processing & Management* Vol. 28, No. 4, pp. 451–455, 1992
Printed in Great Britain.

0306-4573/92 $5.00 + .00

***OPINION PAPER***

# ON THE DIFFICULTIES OF APPLYING THE RESULTS OF INFORMATION RETRIEVAL RESEARCH TO AID IN THE SEARCHING OF LARGE SCIENTIFIC DATABASES

ROBERT LEDWITH
Chemical Abstracts Service, Columbus, OH 43210, U.S.A.

**Abstract**—Although much effort has been applied by researchers to the problem of improving information retrieval systems during the last 20 years, the results of these efforts are not always directly applicable to commercial online systems, especially information retrieval (IR) from large scientific databases. In this paper, the difficulties of extrapolating from the results of IR research to the searching of scientific files accessible via STN International® are discussed and suggestions for further investigation are given.

## INTRODUCTION

When I was asked to contribute an article to this special issue, the editor explained that it would be valuable to have the viewpoint of someone who develops and uses a commercial online system. As a research scientist at Chemical Abstracts Service (CAS), a division of the American Chemical Society (ACS), one of my responsibilities is to evaluate advances in IR for possible application to an online information service, STN International. I have often been concerned about a number of significant differences between the commercial online environment and the typical IR research environment. As a result of contemplating these concerns, I believe the most valuable contribution I can make to this special issue is to discuss the difficulties of extrapolating the results of information retrieval experiments to the problem of searching large scientific databases, and to provide insight on how an online vendor examines advances in information retrieval. Although the article constitutes my opinion, and does not represent the official position of CAS or the ACS, I hope that by describing the difficulties and stating my concerns, each will eventually be resolved.

To help explain some of the difficulties in applying IR research results, this article uses two large scientific files used in a commercial online system and compares them with a large test collection used in IR research, the INSPEC 12,684 collection (Fox, 1983). It briefly discusses differences in the data searched, the searching mechanisms, and the users of the search systems. Using this information, the article discusses how an online vendor wishing to improve service for its current users evaluates online system enhancements. Ranked retrieval methods (one area of IR research) and examples of the differences between the research experiments performed and the current online system are mentioned. I argue that the differences preclude an adequate understanding of the actual performance that a ranked retrieval facility would achieve within the commercial online system. The article concludes with a list of suggestions that could help us acquire a better ability to predict the utility of implementing IR advances within commercial online systems.

## THE DATA

STN International provides access to scientific and engineering information. To help illustrate specific points within the article, two STN files and a test collection used in IR research are used as examples. The first file is the Chemical Journals of the American

Requests for reprints should be sent to the author at Chemical Abstracts Service, P.O. Box 3012, Columbus, OH 43210, U.S.A.

EXHIBIT 2028
*Facebook, Inc. et al.*

Chemical Society (CJACS) File, a primary literature file that contains the text of 97,000 research articles. The second file, the Chemical Abstracts (CA) File, is a secondary literature file that has 9.5 million citations, containing titles, abstracts, keywords, and articulated indexing phrases. These files are compared to one of the larger test collections commonly used in IR research, the INSPEC 12,684 collection. This collection consists of 12,684 document titles and abstracts from the INSPEC database, and 77 queries collected at Cornell and Syracuse universities. (Both natural language and Boolean logic forms of the queries are available and have been used in experiments.) Descriptive statistics for the files appear in Table 1.

## THE SEARCH SYSTEMS

In the STN online service, access to the CJACS and CA files is provided via a conventional Boolean search system that supports the Boolean operators AND, OR, and NOT. (Parentheses may be used to nest expressions.) The documents are divided into fields, such as author name and abstract fields, and search terms may be qualified to match only the occurrences of terms appearing within specific fields. The service also supports searching via proximity operators, which match only the occurrences of specified terms that appear adjacently, or appear within the same sentence, paragraph, or section of a document. When using proximity operators, the user may specify that a variable number of words, sentences, paragraphs, or document sections may appear between the terms being matched. The documents retrieved may be displayed entirely or limited to specific fields.

In research systems, a variety of retrieval models have been used, including the Vector Space, Probabilistic, Fuzzy Sets, and p-norm models. I focused on the p-norm model described by Fox (1983) and Salton *et al.* (1983), which performed well when used to search test collections. In the experiments using the p-norm model, the queries were augmented Boolean queries containing AND and OR operators. Parentheses were used to nest expressions and field-specific queries were supported. Proximity operators were not used in the experiments.

## THE SEARCHERS

Most STN searchers are highly trained in both the domain area and the use of online systems, more so than searchers in information retrieval experiments. Typical STN users have at least one degree in, for example, biology, chemistry, or library science. The users have had several hours of formal training on using the online service and the specific files being searched. Most have refined their searching skills by using the online service for many hours. In short, although some STN searchers are end-users of the data retrieved, the majority are highly trained search intermediaries. In informal discussions, STN users consistently indicate that they are comfortable with the search command language and that they understand and regularly use Boolean and proximity operators. The users are highly motivated to use the system because it is a cost-effective way to find information for a variety of reasons, from preliminary background and SDI searches (typically searches with very low recall and high precision) to patent searches (typically searches with near exhaustive

Table 1. Characteristics of two large scientific files and an IR research collection[a]

| File name | Record type | Records | Term type | Total terms | Distinct terms | Total terms/record | Distinct terms/record |
|---|---|---|---|---|---|---|---|
| CJACS | primary | 96,900 | words | 270,000,000 | 5,536,000 | 2786 | 768 |
| CA | secondary | 9,528,000 | words | 1,234,000,000 | 17,540,000 | 129 | 58 |
| INSPEC 12684 | secondary | 12,684 | stems | 733,800 | 14,683 | 58 | 33 |

[a] For the CJACS and CA files, the words described are limited to those in the title, body, keywords, indexes, and figure titles. Patent numbers, file keys, etc., are excluded, as are the non-searchable words (stopwords). For the INSPEC 12684 collection, the stems are limited to those found in the title and abstracts after removing stopwords.

recall and lower precision). The wide range of search types is different from many of the IR test collection queries, which would be considered as preliminary background or SDI searches by the users of STN.

## EVALUATING ONLINE SYSTEM ENHANCEMENTS

When a self-supporting service such as STN International evaluates a new approach or technology, two primary questions need to be answered:

- To what degree will this benefit the user?
- Is the cost of implementing and using the technology recoverable?

Ultimately, it is the user's perceived needs and willingness to pay for new capabilities that dictates STN's system enhancements. The sci-tech online industry is a small, modest-growth industry, with most online services operating at only a small profit margin. Because there are limited resources for implementing new features, potential online system enhancements must be critically evaluated before implementation. Consider a traditional problem of using Boolean operators. Certain classes of online users have difficulty understanding the function of the Boolean AND and OR operators. As a consequence of this, various schemes involving free-form or menu-based input have been proposed to surmount this problem. STN users, however, have stated that this is not a problem for them; thus the benefit to the current users is minimal. Accordingly, implementing these advancements to the system would receive a low priority.

### *Evaluating the applicability of ranked retrieval to searching large scientific files*

Ranked retrieval models have been examined as alternatives to the standard Boolean retrieval model. However, despite the significant efforts to explore and develop these models, there remain concerns about the models' utility for the searching of large scientific databases. Using the p-norm retrieval experiment described in Fox (1983) as an example, I will present my three major concerns.

1. The first concern is with the size and composition of the collections used for testing in research. Most testing has used small collections containing fewer than 10,000 records or collections containing very brief document surrogates, such as document titles. Of the existing test collections used in IR research, the INSPEC collection, which is one of the larger test collections available, appear to be an appropriate collection for basing extrapolations to the searching of STN files, because it contains both titles and abstracts describing scientific articles. Despite these features, the reliability of extrapolating the performance of research systems that use the collection to a system to search a file over 750 times larger than the collection is highly questionable. At least two factors aggravate any attempts at extrapolation. The first is that a retrieval system must include a human component. Although it is possible to build larger, faster software and hardware components to handle larger files, the human component of the system does not change. In particular, the human cannot and should not be required to review and summarize more data from the larger system than from the smaller one. The second factor deals with the likelihood of unexpected (and undesirable) combinations of terms appearing within the documents, where unexpected combinations cause nonrelevant documents to be ranked as highly relevant ones. To illustrate why this is a potential problem, assume that for a specific set of queries an undesirable combination of terms appears within only .003% of the documents in a collection. If the collection contains 12,684 documents, this is equivalent to one document for every three searches that the user ignores. This occasional document is statistically so small that its influence is easily ignored when examining test results. However, if the collection contains 9.5 million documents, the user must attempt to cope with 285 unwanted documents for each search. Obviously, even very subtle factors within test collection searching could translate into significant effects when searching large files.

2. A second concern is with the nature of the queries used in research collections. Compared to STN user queries, the research queries are too broad. Looking at the INSPEC

collection, a typical query maps to 33 relevant documents out of a collection of 12,684. This would extrapolate to an STN user retrieving and reviewing over 24,000 documents from the CA File. However, a typical STN user reviews fewer than 50 documents per search. Thus, it can be argued that many of the research queries are fundamentally different from STN queries. Another difference between the queries is that most research queries do not use proximity information. This differs from STN user queries, where over 85% of the CA File and virtually all of the CJACS File searches contain one or more proximity operators. The importance of proximity operators may be illustrated by using the example of a user wishing to retrieve information about vitamin A. If one searches for "vitamin" or "vitamins" and "A" in the CA File, 45,800 records are retrieved. However, requiring that "A" must immediately follow "vitamin" or "vitamins" causes only 21% of the records from the first search (9,950 records) to be retrieved. For the CJACS File, the results are even more extreme, with 1500 records retrieved for the first search and 13% of the records (190 records) retrieved for the second. Clearly, using proximity operators can be a valuable tool for improving the precision of some searches of large files.

3. The third concern deals with the performance of ranked retrieval systems and the perceived benefit versus cost to the user. Ranked retrieval schemes are intrinsically more expensive to perform than the unranked schemes. For users to be willing to pay substantially more for a service, they must perceive a noticeable and valuable improvement. However, there are concerns about whether the performance of the existing ranked retrieval models is a large enough improvement over the Boolean search model to represent a cost-effective alternative. To illustrate, assume that a standard Boolean search retrieves 100 documents from a collection, of which 10 are relevant. To find the 10 relevant documents, the user might review 90 documents. A ranked retrieval search such as the p-norm model might also retrieve the same 100 documents, but orders them in an attempt to place the relevant documents first. To find all ten relevant documents, the user might review only 70 documents. While this is a statistically significant improvement in retrieval, in the eyes of the user, it may not be worth the additional cost.

## SUGGESTIONS FOR FURTHER RESEARCH

Having raised these concerns, what suggestions can be made to resolve them? From the perspective of an online vendor of large scientific databases, there are several suggestions:

1. *Research collections with larger vocabularies and more records are needed.* For testing the retrieval of primary and secondary literature, the collections must be large enough to capture the size and complexity of the files that the collections represent.

2. *Investigations of retrieval schemes that incorporate proximity information are needed.* As was shown in the vitamin A example, when larger collections are searched, proximity information may be a valuable aid for improving precision.

3. *Test collections that contain more specific queries are needed.* If large research collections become available, it will be possible to conduct meaningful experiments using queries that correspond to minute portions of collections' records. This will permit better modeling of the types of user searches than is possible with the existing collections.

4. *Investigations into how the human component of the search system can be made more tolerable are needed.* As illustrated in an earlier example, even a statistically small percentage of nonrelevant documents may translate into an unacceptable number of records for the searcher to cope with. Possible mechanisms that might assist the user include aids to integrate, summarize, and display search results.

5. *Investigations into retrieval schemes and search languages for accessing primary literature are needed.* Specifically, the creation of new operators other than Boolean and proximity operators could potentially be very valuable. As studies (such as Ro, 1988) have shown, when searching primary and secondary literature files that represent the same documents, the precision level of the primary literature search is usually much lower than the equivalent secondary literature search. This drop in precision combined with the increased size of primary literature records over secondary ones implies that the searcher's need for

improved access to and concise display of primary literature is even more crucial than when searching secondary literature.

## CONCLUSION

Although it is difficult to determine whether some IR research results may be meaningfully applied to searching large scientific databases, there are efforts underway that recognize the gaps between traditional research efforts and commercial systems. Three such efforts are (a) a proposed investigation into the effect of proximity by Keen (1991); (b) an exploration into issues dealing with a large online collection of chemical primary literature articles within the Chemical Online Retrieval Experiment (CORE) research project (the project is a collaborative effort of OCLC, ACS, CAS, Bell Communications Research (Bellcore) and the Albert R. Mann Library at Cornell); and (c) the development of concept-oriented databases for IR as an alternative approach to searching existing large text databases (Ledwith, 1988). Efforts such as these could eventually lead to resolving the concerns that I have discussed.

## REFERENCES

Fox, E. (1983). *Extending the Boolean and vector space models of information retrieval with p-norm queries and multiple concept types.* Unpublished doctoral dissertation, Cornell University, Ithaca, NY, USA.

Keen, E.M. (1991). The use of term position devices in ranked output experiments. *Journal of Documentation, 47*(1), 1–22.

Ledwith, R.H. (1988). Development of a large, concept-oriented database for information retrieval. Paper presented at ACM Conference on Research and Development in Information Retrieval, Grenoble, France.

Ro, J.S. (1988). An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval. I. On the effectiveness of full-text retrieval. *Journal of the ASIS, 39*(2), 73–78.

Salton, G., Fox, E., & Wu, H. (1983). Extended Boolean information retrieval. *Communications of the ACM, 26*(12), 1022–1036.