SDI Obtained from Numerous Tape Services

Björn V. Tell
Royal Institute of Technology, Sweden
SDI OBTAINED FROM NUMEROUS TAPE SERVICES

by

B. V. Tell.

Abstract

The information service performed by the Royal Institute of Technology handles requests from both university users and from industry. Fourteen data bases are currently used for SDI purposes. As the library also has on-line facilities for retrospective searches of six data bases in the ESRO/RECON network, it has been found that retro-searches usually initiate SDI-profiles for current awareness information.

In 1967 the Institute initiated its own data base "Mechanical Engineering - MECHEN" and began a subscription to the tapes from the Institute for Scientific Information. The purpose was primarily to provide a service to Swedish manufacturers in the mechanical engineering field. However, the interdisciplinary nature of the ISI tapes attracted interest from many users in the universities. The total number of profiles in 1972 was 1,050. The Institute welcomes profiles from outside universities or companies through the intermediary of their own librarians. However, most contacts are carried out directly with the users. The back-up service giving full documentation is necessary for keeping the interest of the users, and places a heavy burden on the regular library service.

This report deals with the coverage, the operational and the performance functions. The system is of a general nature which permits inclusion of various data bases of completely different tape formats. Any combination of elements of a bibliographic record can be searched. As the profile maintenance program becomes an essential element when the number of profiles is building up, an on-line facility for updating has been installed. Thus, it is possible from a type-writer terminal to initiate, update and revise a profile at any time. Two such remote terminals are in operation, one 60 miles from Stockholm.

Special emphasis is laid upon the principles for free text searching and the problems of arranging the printout in such a way that it gives user satisfaction. At present a statistical approach is in operation. Since many data bases have keywords or other subject indicators, a combination of free text search of titles and keywords is often used. There is no significant deviation in the user's reaction when subjected to printout from free text searches or from keyword searches.
Introduction

Within the overriding theme for this IATUL-seminar on "Computer-Based Information Services" my role is to deal with SDI; especially when it is supplied from a number of tape services. Perhaps I should indicate as a start that I am unable to accept that SDI is something apart from other computer-based information services. Since our centre has become a node in the ESRO and LIBRIS network, there is no longer a clearcut distinction between an SDI-service and retrospective searches, nor between on-line and batch processing.

Although I shall focus interest on SDI service, much of what I say could also be related to retrospective searches. The on-line terminal in our centre to the ESRO data bank in Darmstadt, holding 1.2 mil. references, often serves as an entry point for a retro-search for a question which is then placed as an SDI-profile in batch processing mode; here the same number, or sometimes more, data bases are used for giving an exhaustive answer. In total we run fourteen data bases for SDI purposes.

During the years several tape services have been tested and some have been put into operation. Various principles have governed the search procedures, and the output presentation. I propose, therefore, a tripartite examination of the functions of the SDI service in its relationship with 1) the funding agency which initially had to bear the costs as tape manufacturers often have high price structures, 2) the operational centre as having the responsibility for giving a good service, and 3) the users upon whose satisfaction the service depends.

The main points are:

a) The coverage function, the comprehensiveness of the service
b) The operational function
c) The performance function

These functions are not mutually exclusive; there is considerable overlap between what I call the operational and the performance function, but they will help us to understand what an SDI system can achieve.

The coverage function

In the seven years the SDI service has been operating, a number of tape services have been examined, tested and introduced. The funding agency which initially sponsored and supported this new type of service had especially in mind a service to industry, and talks began with the Federation of Swedish Industries about the coverage the SDI service should embrace.

It seemed logical to start off with an interdisciplinary data base, and the choice fell on the Science Citation Index tapes, which also had the advantage of being rapidly produced, so that we could cope with the demand for novelty. Its broad coverage, especially by the citations, assured a serendipity factor of a certain magnitude for those who believe in cross-fertilization between disciplines.

Nevertheless one would not regard this data base as being sufficient for industry, and even if the Institute for Scientific Information from time to time included in the data base new journals from the technological field, the coverage was too meagre to satisfy most of the users from industry. On the other hand, scientists in the university area were extremely pleased with the service when their field of interest corresponded with the ISI coverage.
Obviously, the gaping void especially in the mechanical engineering field had to be filled, and hence the reason for starting up the MECHEN data base. Some industries were interviewed about their needs to cover certain journals. As engineers and technicians in industry are conversant with two or three foreign languages, it was apparent that not only English language journals, but also those in German and French would have to be covered. Later, when we initiated the WOOD data base, Scandinavian journals also had to be included, but references to these were translated into English. We found that with the equivalent of two full time clerical assistants we had the capacity to cover around 200 journals. As small notes, reviews and book critics seemed to be of interest to industry, references to these were also included in the data base.

The creation of the data bank gave us valuable experiences in string handling technique and how to build a general tape format, which will be dealt with later on. It also became a powerful argument against any overpriced commercially available tape service, and it served as an asset for bargaining with government-created data bases in other countries. The production cost could be estimated from the 60,000 references we put, annually, on magnetic tape.

MECHEN and WOOD certainly filled gaps for the mechanical engineering industry and for the paper and pulp trade. When the computerised version of Engineering Index came on the market under the name of COMPENDEX, we included it in our service, despite its badly spelled indexing terms and high error rate. One important sector of Swedish industry is the electrotechnical and electronic field, whose demands were not satisfied until we received INSPEC. This was done on a bargaining basis, not by offering our data base but by giving our ABACUS-program to the Institution of Electrical Engineers, which was then used for many years for its SDI service in Great Britain.

When the particular features of the acquired data bases had become clear to many of the users, the demands for gap filling became more acute. The characteristics which guided our further choice have been journal coverage, speediness and the price structure. We have not bothered very much about the indexing procedure, which will be dealt with later on, or whether the data base contained abstracts or not. Basically a reasonable coverage has been our main goal, because from the beginning our policy has been to answer the question of the user in the best way, disregarding how many data bases are required. This policy has not prompted us especially to advertise any specific data base as particularly valuable for a user. Instead, Bradford's law has been amply verified every time a user thinks he knows the suitable data base for his query. The most pertinent references often stem from unexpected sources. The number of data bases used for different profiles are shown in Table 1.

The coverage function should certainly be looked at with some care keeping in mind the interest of the users. Thus, ERIC was acquired primarily as a data base which could serve the need of personnel managers in industry and organisers of enterprise training courses. Later it became obvious to us that general educators, teachers, research workers and counsellors were the potential market for ERIC and about 100 profiles from this category were received. However, it was often found that pertinent references for these users were pulled out of data bases like INSPEC, ISI and COMPENDEX (Tell, Wessgren, Hemborg 72).

The needs of the present user population are covered by data bases to such an extent that the very great part feels assured of exhaustiveness in their scientific and
technological fields. On the other hand the demand from industry embraces also to a large extent commercial and economic information, an area where no specific acquisition decisions have yet been made. Test tapes of PREDICAST and EXXON are under study.

The search capability and operational function

Basically the creation of the data base MECHEN grew out of the computerisation of an acquisition list (Tell 68) which contained technical reports and was supplemented by a KWOT-index (Key-Word-Out-Of-Titles). This index brought about a string handling technique which has been further elaborated. The primary goal has been to achieve a general system of great hospitality and flexibility (Tell, Larsson & Lindh 70). Paradoxically, numerical analysis rather than computational linguistics crept rapidly into the systems work which had to pass through various stages of development. By a mere masking-off technique where the profile words were kept in the memory and the string of title words was read into the CPU in order to mask off coincidences, we had come far from just searching keywords in assigned fields, and instead we were able to search words allowing for truncations both from left and right - a useful feature for searching chemical compounds. That technique was used for all letters in the alphabet. Even if we could have improved on that technique by using only low frequency letters, we embarked instead on new search principles, namely to make use of tree-structures which gave a considerable advantage in speed over the masking-off program.

As was shown by our collaborator (Dahl 70) the technique could also be improved by using diagram trees instead of single letter trees. But in parallel we embarked on quite another avenue, namely taking advantage of the hash coding address technique (Murray 68) which brought down the search of a title string to milliseconds.

The present system, VIRA, written by Rolf Larsson, is a symbiosis of the tree structure technique and the hash coding in fact means that with regard to CPU time the search procedure is almost negligible compared with the print-out time (Zennaki 72), even for tens of thousands of references matched with an equal number of search terms.

The profile editing program for organising the query terms and the search logic has also gone through various development stages, providing for search on any data base, in any field, and using Boolean or simple arithmetical interpretation of the search question.

Various weighting schemes have been applied. The last year has been one of strong development of an on-line profile program using a type-writer terminal and a dedicated disk of IBM 360-75 for batch processing as a transitory solution. Later a fully on-line dialogue profile program will be used.

Operating an SDI service under a research-granting council means that our reporting and research applications must always expect a more thorough scrutiny of their novelty and experimental quality than would be the case for an activity within the regular budget. This has prompted us to seek advice and carry on a dialogue with interested parties. Sometimes the nature of the work has been such that we have found very few knowledgable people in the field with whom a dialogue might be fruitful. This has often been the case in the systems development field described above, and therefore the scientific literature has been the main source of information. On the other hand, in the development of the profile editing program there has frequently been ground for wider discussions leading to constructive work for improving the service.
Table 1. Number of multiple locations of 1050 SDI queries

<table>
<thead>
<tr>
<th>Data Base</th>
<th>No. of SDI Profiles</th>
<th>No. of Group Profiles</th>
<th>Total No. of Profiles on the Data Base</th>
<th>Data Base Usage Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISI</td>
<td>840</td>
<td>21</td>
<td>861</td>
<td>0.50</td>
</tr>
<tr>
<td>MECHEN</td>
<td>380</td>
<td>17</td>
<td>397</td>
<td>1.53</td>
</tr>
<tr>
<td>CAC</td>
<td>344</td>
<td>3</td>
<td>347</td>
<td>0.34</td>
</tr>
<tr>
<td>INSPEC</td>
<td>552</td>
<td>3</td>
<td>555</td>
<td>1.33</td>
</tr>
<tr>
<td>METADEX</td>
<td>295</td>
<td>7</td>
<td>302</td>
<td>1.92</td>
</tr>
<tr>
<td>NYFLI</td>
<td>237</td>
<td>-</td>
<td>237</td>
<td>0.38</td>
</tr>
<tr>
<td>NSA</td>
<td>52</td>
<td>-</td>
<td>52</td>
<td>0.30</td>
</tr>
<tr>
<td>COMPENDEX</td>
<td>718</td>
<td>19</td>
<td>737</td>
<td>1.60</td>
</tr>
<tr>
<td>ABIIPC</td>
<td>62</td>
<td>-</td>
<td>62</td>
<td>1.07</td>
</tr>
<tr>
<td>WOOD</td>
<td>39</td>
<td>-</td>
<td>39</td>
<td>1.10</td>
</tr>
<tr>
<td>ERIC</td>
<td>131</td>
<td>1</td>
<td>132</td>
<td>0.96</td>
</tr>
<tr>
<td>STAR &amp; IAA</td>
<td>7</td>
<td>-</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>FSTA</td>
<td>54</td>
<td>1</td>
<td>55</td>
<td>0.74</td>
</tr>
<tr>
<td>STU</td>
<td>1108</td>
<td>22</td>
<td>1130</td>
<td>2.92</td>
</tr>
</tbody>
</table>

4819  94  4913

Remark: The Data Base Usage Factor has been calculated as the ratio of the number of references listed in the users' printout and the number of references in the data base which has been processed. It shows that specialised data bases like METADEX or STU (Research projects sponsored by the Swedish Board for Technical Development) have higher usage factors than bases of general nature like ISI. The figures have been calculated on a total of 1,003,800 references.
The main reason for using the representations of scientific papers in the form of titles and using natural language searches, is that if the representation matches the one expressed in the user's profile it is assumed that they are related. The matching does not need to be complete (Simon 69, Zadeh 72). Even redundant matching can, from an operational point of view, be regarded as giving a set of alternatives in which a more correct matching might be embedded. A continuous dialogue with the user and a revision of the profile takes place as a routine matter. The introduction of an on-line profile updating technique has proved especially useful. By feedback from the user, learning might later be embodied in the system (Heaps 70) which will facilitate future matching.

In order to introduce into the system an ability to arrive at gradually more correct matching, a statistical approach has been initiated that will serve to screen some of the alternatives (Sparck Jones 71).

If the system has discovered a match, even only a partial one it may proceed on the assumption that it has detected a correct match whenever such partial matching is present, and go into further stages of refinement by using the weighting procedure.

What happens is that we want on the one hand to equate the selectivity with some kind of feedback of information from the environment, i.e. the reactions of the user to an early output, and on the other we want to check the selectivity against the information inherent in a larger set of the data base than just the subset at hand for the SDI search, and frequency statistics of word occurrence give a guidance for this purpose.

However, we are aware of the fact that we live in a highly redundant world, where the number of coincidences which can be related to statistics in such a way that the statistics really prove effective might seldom occur, still more so if we also take into account compound word expressions. But it would be unwise to ignore this facility which now is built into the system.

So, for instance, a neologism would hardly ever appear in the frequency listing of a large set of references built up from the previously received tapes of a data base. However, as soon as a user is aware of a new word, it is entered into his profile, and can prove to be effective when searching new tapes. Other document representations like keywords or subject categories require a time-consuming intellectual translation process before the authority file can be decided upon, and this puts constraints upon the rapidity with which such neologisms can be picked up. This, in turn, negates to a certain extent the advantage of using fast computer processing in the first place.

Our experiments with the statistical approach have led to useful by-products. The ability of the system to produce frequency lists of words used in titles has given us the opportunity to submit such a list for all the ISO standards to the ISO/INFCO Group when constructing the ISO Thesaurus. The Council of Europe has also requested a frequency list of all title words used in 50,000 articles and reports in the ERIC data base. That list will serve as a basis for the construction of the EUDISED Thesaurus (DECS/DOC 72/15).

Hopefully, the natural language approach can also contribute to knowledge about how scientists write articles, and serve as a basis for international recommendations about this. To this end I proposed to Unesco/UNISIST that they place a contract about this with an American consultant (UNISIST/V/DC/72/1.2 p. 5: 4d).
The profile editing program is a key issue and on its flexibility depends very much both the updating procedure and the search performance. The on-line updating facility was introduced in response to the need for a more labour saving procedure and immediate contact with the users. Many search options and alternative search strategies have been successively included in the program in order to satisfy a variety of needs. Still more development should go into this area to achieve more adaptability and user friendliness (Gluchowicz 71).

The performance function

Much of our attention has been given to the principles that govern the performance of the system as viewed by the user, namely the evaluation he attributes to the output. The funding agency still regards our system as experimental, and it might be of interest to disclose our present thinking about the system performance from the operational point of view we take.

In general, the first step towards an understanding of the kind of literature that a set contains - as a collection of documents, or a file of references - is to develop a taxonomy. This has been the basis for the traditional usage of classification schemes and later the various kinds of thesaurus approach which are in use. This approach anticipates a more or less static collection from which the taxonomy is developed in retrospect.

However, for an SDI service where each new tape might include new concepts we have furthermore to allow for change. Before the step of developing a taxonomy can be taken, we might instead like to look at the representations of the items in the set per se. Usually the representations consist of titles in natural language. The scientists have many ways of representing their research results in the form of titles. Our knowledge about the different ways in which these can be put is incomplete, and in the present state of the art in semantics we would not try to embark on an avenue which would force us to resolve ambiguities, for instance, by transforming each title into its canonical form. This is because we have much less knowledge about the significance of the differences between the various kinds of representations other than the canonical that might be used to express a title.

Not much thinking has actually gone into the manipulation of representations in the title form, since most interest especially in computational linguistics has concentrated on the processing of full text. Although the theoretical basis is practically non-existent, experiments have been directed towards studies of the behaviour of information retrieval systems, based on title searches by the programming and operation of such systems, whereby the results have been tested on the users.

In 1967 we started studies in this field and we are still pursuing free text searching technique (Tell 72). By constructing profiles from the user's query, expressed in natural language, the profile performs inductions, that is to say, it makes use of the context in which various words occur in titles, and by this it is possible to arrive at an interpretation of the title which might be identical to, or different from, the result achieved by indexing.

The approach is identical to that of a scientist reading a journal contents list, where single words in the title more than the complete context of the words might cause creative associations to arise in his mind. It can be said that the program induces meanings of the searched titles by matching them against the weakly pre-co-ordinated profile words, giving ample freedom for serendipity.
Coming back to the earlier point that in natural language many alternative sentences can be phrased as titles and still mean about the same thing, it is obvious that for questions of an inter-disciplinary nature various data bases must be interrogated. The formulation of the query into a profile must, however, focus on one data base at a time bearing in mind the terminology used in the natural language. It seems therefore, necessary to develop a translation system between various scientific disciplines reflecting the language in the data bases by the generation of vocabularies and concordances for words in natural language (Tell 71). The compilation of word frequency lists which we have carried out for various data bases serves the purpose of giving an understanding about the specific scientific "jargon" used.

We have found that it is possible from the high frequency words to determine the specificity of the "jargon", and each data base definitely has its own "jargon". Thus, if when dealing with natural text there are difficulties when going from one data base to another, we can still use only one profile in free language. On the other hand if we also should deal with the meta-language indicated by the subject terms, descriptor, and thesaurus terms, the translation problems become still more serious, especially since each data base has its own thesaurus.

It is clear that a blending of both the free text searching technique and retrieval based upon indexing seems in some fields to give optimal retrieval performance, while in other fields the results are not conclusive. Both our own experience and that of Harwell in Great Britain has shown that free text retrieval is as efficient as retrieval based on indexing in the nuclear field.

Investigations on the effect of the length of the title on the results of free text retrieval indicate that the optimum is 100-150 characters (Olive & Terry 72), which is why we have found that enriching shorter titles to this range is advantageous in our data base MECHEN.

A study has also been made about the relation between search logic and the title length. The results are inconsistent, but it seems to come out that when the search logic is very strict, i.e. specifying a logical product of three concepts (A&B&C) there seems to be a negative correlation, that is to say, if these concepts occur they should come rather close together and the title should not be too long.

Conclusions

To sum up and add a few comments: I have tried to show that the principle of giving satisfaction to the demands of the users has prompted us to acquire a number of data bases upon which the profiles are searched. This has proved Bradfords' law.

The development of the free text search technique - from a mere masking-off technique to a symbiosis of tree-structures and hash coding - has led to search costs which are more than competitive with those for indexed files. The free text retrieval also has an economic advantage so long as the costs of title enrichment are less than indexing costs, and the retrieval results hitherto seem to be equally good. The profiling costs for a free text profile are lower for a profile which has to be searched on several data bases than for a profile based upon the thesaurus language of each data base.

The fuzziness of the natural language approach is improved by the use of a weighting procedure based upon the word frequencies inherent in each data base.

Reactions do not deviate significantly when the user is exposed to references pulled out by natural text searches as compared with searches on index terms.
References

Tell, B.V., Wessgren, K. & Hemborg, W.,
The use of ERIC tapes in Scandinavia, searching with thesaurus terms and natural language.

Tell, B.V.
ABACUS-AB Atomic Energy Computerised User-Oriented Services: The mechanisation of bibliographic list production.

Tell, B.V., Larsson, R., & Lindh, R.,
Information retrieval with the ABACUS program - an experiment in compatibility.
IAG J 1970 (3)4 p323-41.

Dahl, R.,
Fast algorithms for natural text searches in SDI systems.

Murray, D.M.
A scatter storage scheme for dictionary lookups.
Ithaca, Cornell 1968.

Zennaki, M.,

Gluchowicz, Z.,

Tell, B.V.,
Free text retrieval and an ordering for the printout.

Simon, H.A.,
The science of the artificial.

Zadeh, L.A.,
Heaps, H.S. & Ko, W.C.C.,
Automatic adaptive processing of questions in document retrieval.

Sparck Jones, K.,
Automatic keyword classification for information retrieval.

Tell, B.V.,

Olive, G. & Terry, J.,
Title length vs. retrieval results using index terms and natural languages of titles. Harwell, AERE 1972.

DISCUSSION

Z. KOVATS: Could Dr. Tell say whether a catalogue of various tapes exists?

B.V. TELL: There are various catalogues containing descriptions of tape services; one contains around 400 tapes (Directory of Computerised Information for Science and Technology. Ed. by Leonad Cohan. New York, Science Associates 1968 - or Survey of Science and Technology Tape Services. Comp. by Kenneth D. Caroll, New York, Am. Inst. of Physics 1970. - AIP-ID-70-3 or ASIS SIG/SDI 2) and copies are available for sale.

I. BERG HANSEN: Is a regular service provided from ISI Citation tapes?

B.V. TELL: On the whole we have found that using the ISI Citation tapes give too much noise and that it is better to use the ISI Source Index tapes for SDI purpose. What we do with the ISI Citation tapes is that we try to calculate frequency distributions of cited authors and cited articles which is a big job, namely to handle 600,000 citations a year in the computer. We use the ISI tapes to find in which direction a certain type of research is going, or whether there has been an increase/decrease in the citation rate of certain authors.

I. BERG HANSEN: With regard to procedures for automatic updating of profiles, how do users react to changes in the weighting of terms? Would Dr. Tell also give us a little more information on the automatic ranking of output?

B.V. TELL: We do not have automatic updating of profiles, what we have is an on-line updating procedure. Any change in the profile is decided by the user rather than the service. The ranking of the references according to the automatic weighting procedure is in an experimental stage, and only few people are involved in evaluating the weighting. It is difficult to carry out controlled experiments, since the observer is himself too much involved. For example we have found that a particular citation may be regarded as unimportant when it is first seen, but after exposure to a regular service the observer may change his opinion if the same citation is presented a second time. It is not possible to generalise from the results.
I. BERG HANSEN: Would Dr. Tell tell us how MECHEN is evaluated and the cost of running the service?

B.V. TELL: The service is monitored by the number of requests for photocopies. MECHEN includes small notices and news items and there is a demand for this material. The service costs must cover the salaries of two full time clerical assistants and the paper tape.

E. TORNULD: It would be interesting to know how the group profiles put out by the Royal Institute of Technology are received by the users. Is there a temptation to save subscription costs resulting in user dissatisfaction?

B.V. TELL: Standard profiles are appreciated by some customers who do not expect too much from the service. We have, for example a water pollution profile which has 120 customers largely from the paper trade. Standard profiles may often lead to a subscription to the full-fledged SDI service. The functions of the two types of service can be compared to those of a Volkswagen and a Cadillac respectively.

R.A. WALL: May we have more details on the hash coding techniques? The later reference to 16-character strings would seem to cause difficulty with French and German references (which are not translated like those of WOOD).

B.V. TELL: A tree-structure search based on the first 3 characters is employed first. This gives pointers to all terms beginning with the coinciding characters. The number of characters of each term is also given, and the hash-length depends on that number. Hash coding matches are then applied to the term groups in question. The procedure is, therefore, independent of language problems.