DOCUMENT RESUME

ED 089 657 IR 000 353

AUTHOR	Tell, B.
TITLE	SDI Obtained from Numerous Tape Services. IATUL [International Association of Technical and University Libraries] 5th Triennial Conference, Copenhagen.
INSTITUTION	Royal Inst. of Tech., Stockholm (Sweden).
REPORT NO	TRITA-LIB-1050
PUB DATE	17 May 73
NOTE	13p.
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	*Computational Linguistics; *Data Bases; Industrial Technology; *Information Services; On Line Systems; Performance Factors; *Relevance (Information Retrieval); *Search Strategies; Word Frequency
IDENTIFIERS	Free Text Searching; SDI; *Selective Dissemination of Information; Sweden

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. 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. Special emphasis is placed on the principles for free text searching and on the problems of ordering 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 of the user's reaction when given printout from free text searches or from keyword searches. (Author)

LI

IATUL 5th TRIENNAL CONFERENCE, COPENHAGEN

SDI Obtained from Numerous Tape Services

By B. Tell, Royal Institute of Technology, Stockholm

STOCKHOLM 1973-05-17

SDI Obtained from Numerous Tape Services

By B. V. Tell, Royal Institute of Technology, Stockholm

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 begun a subscription to the tapes from the Institute for Scientific Information. The purpose was primarily to provide a service to the 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 documentations is necessary for keeping the interest of the users, and places a heavy burden on the regular library service.

The 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, up-date 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 to order 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 of 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 "Computor-Based Information Services" my role is to deal with SDI; especially when it is given out from numerous 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 center has become a node in the ESRO and LIBRIS network, there is no longer a clearcut distinction between a 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 center to the ESRO data bank in Darmstadt, holding 1,2 Mil. references, serves often as an entry point for a retro search for a question which then is placed as a SDI-profile in batch processing mode, whereby the same number or more data bases are used for giving an exhaustive answer. In total we run fourteen data bases for SDI purpose.

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 center as having responsibility to give a good service and 3) the users upon whose satisfaction the service depends:

a) Its coverage function, the comprehensiveness of the service
b) Its operational function
c) Its 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 perceive what a SDI system performs.

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 begun 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. That had also the advantage of being rapidly produced, so 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 believed in crossfertilization over the discipline borders.

Yet one would not contemplate this data base as being sufficient for industry, and even if the Institute for Scientific Information by and then included new journals in the data base 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, thus the reason for starting up the MECHEM data base. Some industries were interviewed about their needs to cover certain journals. As engineers and technicians out in industry are good in two-three foreign languages, it became apparent that besides English journals also German and French journals had to be covered in these languages. Later, when we initiated the WOOD data base, also Scandinavian journals had to be included, but references to these were translated into English. We found that with the equivalent to two full time clericals we had the capacity to cover around 200 journals. As small notes, reviews and book critics seemed to be of interest to industry, also references to these were 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 barting with government created data bases in other countries. The production cost could be estimated thoroughly from the 60,000 references we annually put on magnetic tapes.

MECHEM and WOOD certainly filled gaps for the mechanical engineering industry and for the paper and pulp trade. When the computerized 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, which demands were not satisfied until we received INSPEC. This was done on a barting bases, not by offering our data base but by giving over our ABACUS-program

to the Institution of Electrical Engineers, which then for many years served for its SDI service in Great Britain.

Now that the property of the data bases acquired had become visible for many of the users, the demands for gap filling became more acute. The characteristics which guided our further choice has 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 if 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 that 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 organizers of enterprise training courses. Later, it became obvious for us that general educators, teachers, research workers and counsellors were the potential market for ERIC and around hundred 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 present user population is covered with 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 decision has 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 MECHEM grew out of the computerization of an acquisition list (Tell 68) which contained technical reports and was supplemented by a KWOT-index (Key-World-Out-of-Titles). This index brought about a string handling technique which

Table 1. Number of multiple locations of the 1050 SDI queries.

Data Base	No. of SDI Profiles	No. of Group Profiles	Total No. of Profiles on the Data Base	Data Base Usage Factor
ISI	840	21	861	0,50
MECHEN	380	17	397	1,53
CAC	344	3	347	0,34
INSPEC	552	3	555	1,33
METADEX	295	7	302	1,92
NYFLI	237	-	237	0,38
NSA	52	-	52	0,30
COMPENDEX	718	19	737	1,60
ABIPC	62	-	62	1,07
WOOD	39	-	39	1,10
ERIC	131	1	132	0,96
STAR & IAA	7	-	7	-
FSTA	54	1	55	0,74
STU	1108	22	1130	2,92
	4819	94	4913	

Remark: The Data Base Usage Factor has been calculated as the ratio of the number of references listed in the printout to the users and the number of references in the data base which has been processed. It shows that specialized 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.

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 more than computational linguistics crept rapidly into the systems work which had to pass through various stages of development. By a more 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 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) that 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 which in fact means that with regard to CPU time the search procedure is almost negligable even for tens of thousands of references matched with an equal number of search terms, compared with the print out time. (Zennaki 72).

The profile editing program for organizing 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 typewriter 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.

To operate a SDI service under a research granting council requires that our reporting and research applications always have to expect a more thorough scrutinity of their novelty and experimental quality than would be the case for an activity within the regular budget. This has prompted us to seek advise 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, instead 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 a ground for more wide discussions leading to constructive work for improving the service.

The profile editing program is a key issue and on its flexibility depends very much both the updating procedure and the search performance. The introduction of the on-line updating facility responded 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 work 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 are taking.

In general, the first step towards an understanding about what kind of literature there is in a set 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 approaches which are in use. This approach anticipates a more or less static collection wherefrom the taxonomy is developed in retrospect.

However, for a SDI service where each new tape might include new concepts we have furthermore to allow for change. Before the step to develop a taxonomy can be taken, we might instead like to regard the representations of the items in the set per se. Usually, the representations consist of titles in natural language. The scientists have many ways to represent their results of research in the form of titles. Our knowledge about the different ways in which it can be put is incomplete, and at the present state of the art in semantics we will not try to embark on an avenue which would force us to resolve ambiguities,

for instance, by transforming each title into its canonical form, because we have much less knowledge about the significance of the differences between the various kinds of representations other than the canonical that a title might be expressed into.

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 experimenting has gone on in studying 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.

Already 1967 we started studies in this field and we are still persuing free text searching technique (Tell 72). By constructing the profiles out of 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 a kind of interpretation of the title which might be identical or different from what is arrived at when indexing a title. The approach is identical to that of a scientist reading a list of contents of a journal where every single words in the title more than the complete context of the words might arise creative associations in his mind. It can be said that the program induces meanings of the searched titles by matching them against the weakly precoordinated profile words giving ample freedom for serendipity.

The main idea 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 reduntant matching can from an operational point of view be regarded as giving a set of alternatives where 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. Especially the introduction of an on-line profile updating technique has proved 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 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, 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 occurrencies gives a hint for this purpose.

However, we are aware of the fact that we live in a higly redundant world, why the number of coincidences which can be related to the 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 nelogism would hardly or ever appear in the frequency listing of a large set of references built up from earlier 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 which puts constraints upon the rapidity with which such nelogisms can be picked up. That negates to a certain extent the advantage of using fast computer processing in the first place.

Our experiments with the statistical approach has 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. 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 the knowledge about how scientists write titles, and serve as a basis for international recommendations about this. To this end I proposed to Unesco/UNISIST to place a contract about this by an American consultant (UNISIST/V/DC/72/1.2 p. 5: 4d).

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 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 considering 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 has definitely its own "jargon". Thus, if already 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 metalanguage which subject terms, descriptors, thesaurus terms indicate, 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 in some fields seem to give optimal retrieval performance, 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 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), 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

Just 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 the Bradford's law.

The development of the free text search technique - from a mere masking-off technique to a symbiosis of tree-structures and hash coding - have led to search costs which are more than competitive with those for indexed files. The free text retrieval has also 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 is 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.

The user reactions do not deviate significantly when he is exposed to references pulled out by natural text searches compared with searches on index terms.

References

Tell, B.V., Wessgren, K. & Homborg, V., The use of ERIC tapes in Scandinavia, searching with thesaurus terms and natural language. Strasbourg, Council of Europe 1972. (DECS/DOC 72:15)

Tell, B.V., ABACUS - AB Atomic Energy Computerized User-Oriented Services: The mechanization of bibliographic list production. - IEEE Trans. on Eng. Writ. & Speech 1968, 11:2 p. 110-117.

Tell, B.V., Larsson, R. & Lindh, R., Information retrieval with the ABACUS programme - an experiment in compatibility. - IAG J 3(1970) 4 p. 323-41.

Dahl, R., Fast algorithms for natural text searches in SDI systems. Göteborg, Chalmers Institute of Technology 1971. 49 p. (Department of Computer Science. No. 71:02) (In Swedish)

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

Zennaki, M., Exposé VIRA. Paris, CNRS-CDHS 1972. 22 p.

Gluchowicz, Z., Selective dissemination of information - a transdisciplinary information retrieval system at the Royal Institute of Technology, Stockholm. - IAG J (1971) 2 p. 131-48.

Tell, B.V., Free text retrieval and an ordering for the printout. Sthlm 1972. (TRITA-LIB-1037) (Report to the ENEA Tutorial Seminar on "Indexing vs. Free Text Retrieval, Studsvik June 1972.)

Simon, H.A., The science of the artificial. Cambr., Mass. 1969.

Zadeh, L.A., Fuzzy languages and their relation to human and machine intelligence. - Man and Computer. Ed. by M. Marois. Basel, Karger 1972 p. 130-178.

Heaps, H.S. & Ko, W.C.C., Automatic adaptive processing of questions in document retrieval. - Proc. Amer. Soc. for Inform. Sciences 7 (1970) p. 319-21.

Sparck Jones, K., Automatic keyword classification for information retrieval. Lond. 1971. 253 p.

Tell, B.V., Global and long-distance decision making. - In: Global and long-distance decision making, by K. Samuelson, H. Borko, B.V. Tell, P. Nador, R. Dubon, P.E. Irick, P.E. Mongar and H.F. Dammers. Stockholm, FID/TM 1972. 83 p.

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