Problems in the empirical evaluation of Information systems

Robert Fugmann

1 Introduction

The goal of information systems is to create order in a collection of documents so that an information searcher need not scan the entire collection in an attempt to find information of interest. As a consequence of the steadily increasing demands made on our information systems, they have been subjected to the necessity of being incessantly developed into higher and higher levels of advancement.

Development work must be oriented towards predetermined goals. The extent to which these goals have been approximated must be assessed through the application of quality criteria. If wrong goals are established and pursued the true goal can never be achieved. Should an information system's development begin to stagnate or to depart from its goals, a new orientation is necessary. Otherwise, development work would yield increasingly unsatisfactory results.

The kind of philosophy that underlies development work plays an important role for the prospects of this kind of work. An inadequate philosophy may well lead research and development into the wrong direction and render it futile. Information system development does not constitute an exception to this rule. Positivism is an example of such an inadequate philosophy.

2 The dominance of the positivistic view in information science

In the positivistic view, reality of phenomena is only acknowledged if they are observable. This philosophy flourished until the second and third decades of this century. It was abandoned for its obvious inadequacy, at least in the natural sciences. Even the idea of atoms and molecules was rejected by prominent positivists because they could not be observed and realized empirically.

Positivism, especially in its variation of empiricism, would support the medieval statement of the direct relation of weight to speed of freely falling bodies: the heavier the body, the faster it falls. Ample empirical evidence for this thesis could be compiled by comparing the free fall of various stones and pieces of paper. Galileo stated the opposite relationship, namely, the independence of falling speed from weight and could also present confirming observations in the case of massive falling bodies.

Mere reasoning would have corroborated Galileo and refuted the opposite statement: if two stones are tied together (and the weight of the body is thus increased) there is no reason that they should fall faster than when they remain isolated. Which of them would have to fall faster under these circumstances and pull the other one and why should this be the case? Obviously, something is wrong with the empirical interpretation of experiments with light bodies. Today we know that air resistance is the factor that disturbed Galileo's experiments. In the vacuum, i.e. after the elimination of air resistance, the speed of falling bodies is independent of their weight.

Empiricism leads us astray when we select the wrong experimental conditions or restrict ourselves to a one-sided selection of examples. Using the wrong instruments is also a common source of fallacy, namely, using only those that merely happen to be available (instrumentalism, cf. also Budd 1995, p. 300). Ample empirical "evidence" could also be supplied for a postulated relationship between a purse's weight and the amount of money it contains if only those conditions are chosen which are conducive to the statement.

In the information field we encounter an instructive example of the adherence to the empiricist view which has led to treacherous results: purely empirically, an inverse relationship between precision and recall seems to prevail in information retrieval. According to this "empirical law", precision cannot be improved without simultaneously impairing recall and the reverse was also observed to prevail. For decades this statement has constituted an excuse for inadequate information systems and has thus paralyzed their improvement, although this type of relationship has often been put in doubt (for example by Soergel 1985, 122; Harter 1990, pp. 136, 145; Green 1992, 87; Fugmann 1994, p. 154; Svenonius 1995, 247).

What is neglected here is that information systems are observed which work perfectly both at 100 % precision and 100 % recall, as is the case, for example, for the majority of chemical molecular structure searches and for searches in the telephone directory.

It is true that an inverse relation between precision and recall is often observed but only under circumstances in which there is no predictability in essence selection and in essence representation during indexing. Under these unfortunate circumstances, one has to find one's search statements by way of trial and error. This lack of predictability occurs

- if there is no index language in use for general concepts and/or
- if an index language is used in an unreliable manner.

The postulated "law" of the inverse relation has been stated in the early Cranfield experiments and has persisted until today.

In the information field and in a purely empirical approach, literally any (intended) evaluation result for an information system can be produced and any opinion on the effectiveness of a type of information system can be corroborated through the choice of or through the adherence to a specific constellation of experimental conditions and, in fact, in a manner which seems perfectly convincing and unsuspicious, just as is the case in the experimentation with falling bodies, if there is no critical contemplation of the experimental conditions.

Empiricism relies on the observable (and on its measurement) and neglects what cannot be observed or is excluded from observation, (cf. Hjorland 1997, p. 59). Hence, empirical "evidence" cannot constitute genuine evidence. But empiricism can, by providing facts, give an incentive to a more advanced interpretation of reality, such that observed phenomena can be better explained and (yet) unobserved ones can be better predicted.

3 The inadequacy of consistency as a criterion of indexing quality

Traditionally, indexing consistency is recognized as a criterion of indexing quality, although this criterion has repeatedly been put in doubt (cf., for example, Cooper 1969; Soergel 1994, p. 594).

This statement neglects the fact that a mode of perfectly consistent indexing may well be consistently defective. An optimum of indexing consistency could easily be achieved through the application of automatic indexing in any variation, for example through merely mechanically extracting keywords from natural language texts using a stop word list. But this may well lead to an ineffective or even unsuable information system because it leads to unpredictable expressione for the concepts of interest.

On the other hand, and intuitively, consistency has something to do with indexing quality. If the same document is indexed entirely differently by different indexers (or indexed differently by the same indexer at different times), something seems to go wrong.

The solution to this puzzle is that consistency is paramount only in the first step of indexing, that of essence selection. Here it is identical with the predictability of the selection. In the second step of indexing, that of essence representation, only predictability is required (Fugmann 1993, p. 94-97).

A concept may well have been entered into the search file with different modes of expression, i.e. markedly inconsistently. But if these expressions can be looked up, they are made predictable. These expressions can then be compiled into a set of alternative search statements. For example, a substance may be represented by a variety of expressions, such as "vitamin C", "ascorbic acid", "cantanR", or a paraphrasing expression like "scurvy-preventing substance in vegetables", etc. But if we know these expressions, we can readily retrieve the corresponding documents.

Hence, consistency is an invalid criterion for indexing quality. What is needed, instead, for good retrieval is predictability, both at the stages of essence selection and essence representation.

4 The number of access points as a criterion for expected retrieval success

It has been postulated that the greater the number of "access points" to a document, the better its retrievability will be. If this is true, full text search files should be ideal for high recall values because they provide the maximum number of access points. But we know that full text files are far from being ideal in this respect (cf. for example, Blair 1996, p. 19). The reason is that the words of the searcher for the topic of interest only rarely match the wording of the documents of interest, especially in the case of searches for general concepts and topics.

Only a type of expression that is phrased in a predictable mode can constitute an access point for retrieval, a condition which is not fulfilled in ordinary natural language. Therefore, indexing languages, with their capability of providing predictable concept representations, may well be more specific in retrieval than an even more detailed natural language text, although an indexed text presents substantially fewer "access points" for the mechanized search, especially in the case of searches for general concepts and topics.

5 Survival power: the neglected criterion of information system quality

It is alien to the empiricist view to look into an information system's future because its future cannot be observed and because the system's fate cannot be empirically demonstrated in the present. What is generally preferred is a snapshot-like analysis of the presently prevailing situation because such an analysis is fast and cheap.

However, an information system user should be highly interested in whether an information system is capable of continuing its service in the future. There may be various reasons to abandon an information system in the future after some time of having been practiced in operational use. The demands made on retrieval precision may be steadily growing during the growth of a mechanized information system and the system may fail to provide significantly higher search specificity. Then it begins to produce hundreds or even thousands of irrelevant responses, from which the few relevant ones have to be sifted out through human inspection. This is a procedure which may well develop into requiring an intolerably high expenditure of time and attention.

An information system may also become unusable because its vocabulary has become chaotic in the course of time. This renders the indexing procedure correspondingly unreliable and the search results become increasingly incomplete as a consequence.

This decrease in recall may remain hidden for quite a long time because it is difficult to observe. But when this deficiency becomes apparent some day, perhaps through a correspondingly extended and careful investigation, an information system suddenly loses all the appreciation that it had so far enjoyed.

The notion of information system survival power, although coined early in the history of our field (Harmon 1970), has not become popular in contemporary information science literature, although this is one of the most important quality criteria for operational information systems.

6 The full text information system

The equivalence of full text information systems with intellectually indexed systems has often been stated and empirically "proved", for example in the Cranfield experiments. In fact, an equivalence (or sometimes even a superiority) seems to prevail, but only if any distinction between individual concepts and general concepts is dispensed with and when, due to the smallness of the experimental files, the experimenters can memorize the modes of expression for general concepts happen to be in the file. Under such circumstances a crucial quality criterion, namely, the predictability of the mode of expression of a concept of interest, is not put to the test.

Here we encounter what has been appropriately called the "small system syndrome", that is the phenomenon that small information systems display properties that are fundamentally different from those of large ones ("non-scalability" of information systems) and that the mere growth of a small system may well lead to its decline and eventual break-down (cf. Gey, Dabney 1990).

What is hidden to the empiricist view and what is therefore neglected in many evaluation studies is the fact that a word used in natural language is not intended for use in isolation. It is only in context that a word assumes meaning and importance and, to the reader or to the listener, the context is presented. Hence, in colloquial discourse, any text requires (and subconsciously receives) interpretation (cf., for example, Budd 1995, pp. 307, 308).

- Text interpretation yields .reliable essence recognition (the importance of a concept depends on the context in which it is embedded),
- .word meaning disambiguation (through the knowledgable utilization of the word's context),
- .paraphrase lexicalization (i.e. substituting a paraphrase in storage or retrieval by one of its lexical equivalents in natural or artificial language, such as notations or descriptors. The

problem of expression multiplicity for a concept in natural language is much more than merely one of synonymy),
- ellipses filling (i.e. making explicit what has only been implied in a text and must be inferred from it, cf. For example, Ranganathan 1962, p.129; Green 1992, p.84; Fugmann 1993, pp. 64, 70, 91),
- establishing concept relationships,
- (near) synonym control,
- verbalization of non-textual information,

all of which are typical achievements of good intellectual indexing. Thus, concept representations are made predictable, and concepts are made easily retrievable. However, these achievements are renounced in searching non-interpreted full texts, much to the detriment of search quality.

As far as context is concerned, it is always freely phrased and, hence, expressed in an unpredictable manner. Therefore, it escapes inclusion in the query as a reliable, interpretative statement.

Text interpretation defies mechanization because it is an inherently indeterminate process. Its point of departure is an indeterminate one, namely, the unpredictable mode of expression encountered in natural language. Hence, interpretation proceeds in an unpredictable manner, too. No instructions can therefore be laid down in advance in a program for the satisfactory, mechanized execution of interpretation. These instructions would have to be infinitely numerous.

Full text storage has often been given credit for the specificity which it seems to provide for searches. Here it seems (and occasionally is) superior to indexing and classification, which often lack sufficient specificity. But specificity without predictability is largely useless for searching as is obvious from the foregoing. Using this specificity will yield both false responses (through the unresolved ambiguity of natural language words) and a dramatic loss of information.

7 User evaluation

Any indexing must aim at satisfying its users, not only in the present, but particularly in the future and also under the changing requirements of the future. It is therefore paramount for an information system that the users are involved in the design of the system and in continually adapting the system to unforeseen new requirements.

Specifically, users must render their opinion as to whether the conceptual categories of the system under discussion meet the requirements for searches and whether navigation in the vocabulary is sufficiently easy and fast. They must give feedback if the specificity of the vocabulary should be extended or reduced. They should urge for timeliness of the input and for an appropriate speed in the execution of the searches. The searches should be affordable, the responses should be easily accessible and should display an appropriate ratio of precision and completeness (whereby the latter is difficult to assess and therefore mostly neglected). Users should also care for the maintenance of an appropriate coverage of their literature in their information system. They should express their opinion about the quality of book indexes to publishers in order to make them aware of the necessity of improving them.

But users must also accept

- that not all present-day requirements could have been foreseen in the past and also cannot be foreseen for the time ahead;
- that an information system must display some features the judgement of which goes beyond their competence, in so far as the user is not an information expert. (Some of these features have been discussed in the foregoing.)

A widespread misconception is caused by neglecting the difference between a delegated and a non-delegated search. When an information seeker searches the literature, he or she is accustomed to encountering a great variety of documents, many of which prove to be of interest although they are more or less distant from the goal that the searcher initially had in mind. In one's own self-service search, one has an entirely free hand to gather what raises interest, independent of any predetermined search goal.

Many users expect that a computerized search will ("at least") yield the same fortunate results. But any delegated search requires the definition of what is to be searched. Only that which can defined can be satisfactorily delegated to another person or to a mechanism). The computerized search does not constitute an exception to this rule. What is pertinent in this sense for an individual user may well not be pertinent for the vast majority of the other users, and may even be pertinent for an individual user only at a particular point in time. But the user would be unable to specify in advance which (still unknown) records would be found "interesting" when and if they are encountered.

Such an expectation is unsatisfiable in a system for the delegated information search because it is typically oriented to what has been laid down in advance as the goal of the search. Satisfying such an expectation would require the omniscient, clairvoyant, and perfectly prognostic programer who would have to lay down a priori what can only be done a posteriori, and then even only by the questioner himself or herself, namely to specify in advance what would be of interest if it was encountered in the future.

Furthermore, trying to approximate such an expectation may well lead an information system directly into its decline and eventual break-down in the more or less distant future (cf., for example, ISKO/IE 1992). Too many false, unrequested responses will be retrieved and make the few requested ones unaccessible.

Instead, it is the task of retrieval to exclude those retrieval responses that are merely pertinent because they have not been requested by the user, however useful they may be for an individual user in his momentary, specific situation. Subjective, serendipitous browsing the literature or the results of excessively generalized queries continues to constitute the appropriate way of finding them, even in the era of computerized information retrieval.

Hence, in assessing retrieval quality, users must distinguish relevant responses (i.e. those responses that meet the elements of the search request) from pertinent ones, i.e. those that happen to be of momentary interest to an individual (!) user but which do not match the search request. Otherwise, even the sequence in which the responses are presented may determine whether or not they are rated "relevant".

It is a misconception of "user friendliness" if the goal is to simulate the highly subjective and undefinable process of serendipitous browsing executed by the user. User friendliness should not be permitted to lead into methodological primitivity which, in turn, will probably lead to the eventual

break- down of the information system, much to the disadvantage of the same users who have insisted and relied on it.

Here, the information specialist must be firm strong in contradicting such an ill-considered expectation on the part of the naive user and in contradicting the corresponding criticism of information systems that do not offer the full serendipity of the non-delegated search.

Neural networks have been given credit for allegedly obviating any a priori instructions because they can autonomously "learn". But they can at best learn from existing experience. This is not sufficient for executing satisfactory indexing of what is continuously coming up in the literature and what has to be processed.

8 Conclusion

A markedly positivistic - empiricist philosphy still dominates the field of information system evaluation and, hence, the design of information systems. This has led to the neglect of those important system features which are difficult to observe and to an emphasis on what can be easily observed (and even measured). The persistence of such a view throughout decades of research and development is not surprising as we know from the history of the sciences throughout the centuries. This, however, should not discourage researchers from defining new paradigms. They may well constitute an incentive to substantial advancement.

Bibliography:

Blair, David (1996): STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years After. Journal of the American Society for Information Science 47, No. 1, 4-22.

Budd, John M. (1995): An Epistemological Foundation for Library and Information Science. Library Quarterly 65, 295-318.

Cooper, William S. (1969): Is Interindexer consistency a Hobgoblin? American Documentation, 268-278.

Fugmann, Robert (1985): The Five-Axiom-Theory of Indexing and Information Supply. Journal of the American Society for Information Science 36, 116-129.

Fugmann, Robert (1993: Subject Analysis and Indexing - Theoretical Foundation and Practical Advice. INDEKS Verlag Frankfurt/M.-Germany (now Ergon Verlag, Gromb□**hls**trasse 7, D 97 080 W□**rzburg-Germany). ISBN 3-88672-500-6.**

Fugmann, Robert (1994): Galileo and the Inverse Precision-Recall Relationship: Medieval Attitudes in Modern Information Science. Knowledge Organization 21, No. 3, 153-154.

Gey, Frederic; Dabney, Daniel P. (1990): Letter to the Editor. Journal of the American Society for Information Science 41, 613

Green, Rebecca (1992): The expression of syntagmatic relationships in indexing: Are frame-based index languages the answer? Classification Research for Knowledge Representation and Organization. Proceedings of the 5th International Study Conference on Classification Research, Toronto, Canada, June 24-28, 1991, ISBN 0 444 89 343 1.

Harmon, Glynn (1970): Information Need Transformation During Inquiry: A Reinterpretation of User Relevance. Proceedings of the American Society for Information Science 1970, Vol. 7,

41-43.

Hjorland, Birger (1997): Information Seeking and Subject Representation. Greenwood Press, ISBN 0-313-29893-9.

ISKO/IE (1992): User Evaluation of Information Systems. International Classification 19, 151-152.

Ranganathan, S.R. (1962): Elements of Library Classification. ASIA Publishing House, London.

Soergel, Dagobert (1985): Organizing Information - Principles of Data Base and Retrieval Systems. Academic Press Inc., Harcourt Brace Jovanovich Publishers, New York. ISBN 0-12-654261-9.

Soergel, Dagobert (1994): Indexing and Retrieval Performance: The Logical Evidence. Journal of the American Society for Information Science 45 (8) 589-599.

Svenonius, Elaine (1995): Precoordination or not? Subject Indexing: Principles and Practices in the 90's. Proceedings of the IFLA Satellite Meeting held in Lisbon, Portugal, 17-18 August 1993. ISBN 3-598-11251-3.

1) Axiom of definability: "The compilation of information relevant to a topic can be delegated only to the extent to which an inquirer can define the topic in terms of concepts and concept relations" (Fugmann 1985, p. 118; Fugmann 1993, pp. 41, 45).