Users satisfaction through better indexing

Gholamreza Fadaie Araghi

ABSTRACT

Classification and indexing are two main tools to organize information to serve the users. Information architecture is nothing more than to organize better to achieve this goal. Any user seeks easy access and speed to reach one's information needs. A classifier/indexer must interpret or estimate the users' need in the best possible terms. Ranking algorithms–such as Boolean, Vector, or others–is highly recommended and practiced. Some define Retrieval Strategies as a measure of similarity between a quarry and document. Relevance is a criterion for matching aboutness. Aboutness is a criterion for decision-making. Better indexing, as well as better classification, is a key to reaching the ultimate goal in record management. Some suggestions are made for those who create databases, provide information engines, or manage the information.

KEYWORDS: Indexing, classification, relevance, aboutness, Boolean, Vector, probability, Retrieval

INTRODUCTION

In the information age, due to a lack of time and an abundance of information, users seek the quickest way to get the right answers. Information provider technologists, as well as librarians and other information providers, try to find the best way to satisfy the users. Classification and indexing, as tools for all information providers, are in two ends of one spectrum. They are both invented to serve users through retrieval. When one goes from generality to specialty, the process begins at classification and ends at indexing. Yet there are a lot of questions in this field. In spite the invention and progress of some applicable methods, such as Boolean, Vector, and Fuzzy systems, indexing still has, and will continue to have, high value in this domain. That is, classification, as well as indexing, must promote in such a way to cope with the problems. Finding a new approach or methodology for better indexing is highly recommended. To reach this point, it is also advised that librarians and information providers must keep up with the users' needs and try to estimate them. I am going to present in this paper how to find some ways applicable in both classification and indexing processing that can get better results in retrieval.

Classification and indexing is for retrieval; so from the beginning, the points of view of classifiers/indexers must match the users' needs. Certainly, what Rowley and Farrow [1] emphasized, that "indexing must be tailored to the needs of user," applies to classification, too. The classification algorithm, says Hynek, [2] is evaluated in terms of accuracy and speed. As a matter of fact, classification/indexing is for the sake of the user and his satisfaction. Therefore, the two most important standpoints, which may satisfy the users, are as follows:

- *Easy access*. This feature is very important for the user, because one wants to have ease in accessing the accurate information. After the development of the automatic retrieval system, computer generations progressed quickly toward easy access. Computer languages are rapidly becoming user friendly. Nevertheless, problems remain in using automatic retrieval systems.
- *Speed*. Another important element that satisfies the user is speed. Users in the age of the information infusion are usually in a hurry and cannot spare a lot of time on searching. One wants to get what he/she needs as soon as possible without any trouble. "Finding the relevant document is not enough, it should be retrieved within one acceptable response time."[3]

SOME CRITERIA AND METHODS FOR RETRIEVAL

There are some methods recently developed for retrieval, and researchers are working on this seriously. Nowadays, retrieval as an outstanding goal in electronic information science has led to the birth of a domain called information architecture. However, as Latham [4] notes, it is not anything more than information organization in a broad sense. Both of them have been involved with library and information studies so far. To have an effective approach to user's consent of what one is looking for, the indexer/classifier should obtain the user's satisfaction. In doing the job, the indexer/classifier may have to interpret or estimate the user's information needs in the best possible terms. These terms must be quite familiar to the user. Predicting relevance or non-relevance of documents, say Baeza-Yates and Riberto-Neto,[5] depends on ranking algorithms. They, like many others, group the models in three, which are Boolean, Vector, and Probabilities. Grossman and Frieder[6]propose eight, under the title of Retrieval Strategies. They define Retrieval Strategies as a measure of similarity between a query and a document.[7]. Their
eight strategies are as follows: Vector Space Model; Probabilistic Retrieval Strategies; Inference Networks; Extended Boolean Retrieval; Latent Semantic Indexing; Neural Networks; Genetic Algorithms; and Fuzzy Set Retrieval, but they emphasize that most research literature focuses on four key retrieval strategies: Vector, Probabilistic, Boolean, and Fuzzy-set.[8]

Relevance is a criterion for matching the aboutness. Aboutness is very important, but very difficult to measure. Estimating the aboutness in humanities is even more difficult. Both indexer/classifier and user are\ supposed to estimate the aboutness, but in different ways. Each of them must try to anticipate the semantic aboutness of the meaning of the words and phrases and use it in their decision-making. To measure the relationships between selected documents, says Losee,[9] bring us some aboutness. This aboutness itself is a problem because we do not know the user's preference.[10] Losee defines[11] that this is besides what the indexing system has determined to be the information fragment. He proposes the binary search based on the answer to the previous question. In such a search, he insists, the sequence of the questions is important. The answer to each question excludes a number of irrelevant queries. [12]

Boolean Queries

Boolean is known as the most popular retrieval system. It is famous for its simplicity and clean formalism. Boolean logic is used in most search systems. A weakness of this system [13] is that it excludes items, which partially fulfill the user's need. That is, although it seems to be very easy to use, its disadvantage is that it may extract many relevant items or retrieves too many irrelevant ones.

Vector

The Vector model is based on the arithmetic calculation. It is said to be a kind of expansion of the Boolean model,[14] with which we can use natural language. With Vector model, by weighting the words, the indexer/ classifier tries to increase the probability of finding relevant informative items.

Fuzzy Retrieval

As the Boolean logic system fails to answer the user's need completely and it searches for just yes or no answers, there is high probability that documents with partially related subjects will be excluded. To measure this probability, the fuzzy system is used to answer the need. In

fact, we have two ways in a language system to deal with in the retrieval system. That is, an artificial pre-coordinated language and a natural one, of which the latter is said to be more powerful in retrieval. Natural language, being more related to indexing, gets more aboutness in the fuzzy system.[15] Yager,[16] through the linguistic specification of the interrelationship between the desired attributes, proposes an extension of the fuzzy set method of information retrieval. This model, Losee[17] points out, may have the feature of Boolean, as well as the feature of the probabilistic model. He continues: " In the fuzzy system, for example, the feature dog would have different fuzzy values, for different documents. One book may have .95 value, the other one may have 0.001 estimate about dog."

This may be the same idea, as Svenonious[18] points out, that Wittegenstein argues, that words do not have not fixed boundaries and that some words have more fluid boundaries than others. Consequently, their category membership is not black and white. Fuzzy set theory clarifies the idea and takes away the ambiguity.

INDEXING/CLASSIFICATION

The gateways through which all documents are retrieved, Rowley and Farrow explain[19], are those in which the classifier/indexer is engaged. These are catalogs, bibliographies, indexes and abstracts, record management systems, and network resources, among which, indexing for good retrieval is vital and critical. As long as indexing is the most significant and powerful technique in information retrieval, and taking into view all other recommendations, certain measures must be taken for better retrieval. These are divided into two main categories, internal and external. By internal I mean all those actions which can be taken by the classifier/indexer in a broad sense and by external, those activities which should be done by all those who are engaged in creating databases and information engine providers, especially information managers. The suggestions are as follows:

Internal

Promotion of the Quality of the Indexing System. There should be enough help in the indexing system denoting term provisions. That is, help to clarify which subject entries exist in a given library or database. No doubt, say Baeza-Yaets and Roberto-Neto,[20] indexing is at the core of every modern information retrieval system. Deep indexing for better retrieval, although this is very expensive and time-consuming, must be seen, in a sense, to comply with the user's need as a researcher. Surely, the ease of access from the researcher's point of view may compensate for the cost. Failing to find the needs by not getting the exact help term, one must shift to choosing more general terms. The main point is that sometimes an indexing system fails to represent the collection in a logical way, or it is not capable of defining it. I myself had an experience in the McLennan Library, at McGill University, Canada as follows:

> In a journal search in MUSE (McGill University System Enquiry), I attempted to find journals about classification of knowledge. The result on the screen was: "Error, not found," but there was a note: *classification* 104, *knowledge*, 78, and the stop word *of*, 411975.

This example and many others show that the principle problem of the retrieval process is in the classification and indexing system. Referring to the debate of Bloomfield with Lancaster in his article *Indexing–neglected and poorly understood*, what is quite clear from his argument is the importance of classification/indexing in this area. Bloomfield clarifies that both of them believe in a good indexing system, but in a little different way. Table 1, comparing 7 and 6 criteria of Lancaster and

Bloomfield respectively, illustrates the differences.[21] Some of the points on the left side, according to my comparison, may relate to more than one element in the right side and it is shown by dotted line.

Promoting the Ability of the Classifier/Indexer. The indexer/classifier must be well aware of the rules of indexing. Education, instruction, and workshops may affect job performance. Hynek, describing the importance of classification/indexing, warns [22] of working with the uncontrolled classifier. He proposes *clustering analysis* techniques. By this, he means that a criterion is not a priori known to enable the classifier/indexer to find them. In this situation one may not use the best possible term. Hynek [23] reports of Schapire's Boosting algorithm by which the training data breaks down into subsets where classifiers either succeed or fail to make predictions. He suggests that a decision tree induction algorithm may be practiced and applied to distinguish the cases where the classifier is correct from where s/he is not correct. Training and testing data is suggested to improve the potentiality of the classifier/indexer in order to get better results. He adds, as each classifier has a particular sub-domain, for good results in classification, multiple learned classifiers should combine their knowledge. In the traditional way of classification/ indexing they are left alone and they may do their job in a void and using their own judgments. This may not completely comply with the real needs of users.

TABLE 1. Comparison of Lancaster and Bloomfield Criteria for Indexing

Lancaster	Bloomfield
1-Number of term assigned	1-See and see also references
2-Controlled vocabulary vs. free text index	2-Breadth of vocabulary
3-Size and specialty of vocabulary	3-Use of general and special terms
4-Characteristics of subject matter and its terms	4-Indexing format
5-Indexing factors	5-Depth of indexing
6-Tools available to indexer	6-Inclusion of titles or other subject qualifying phrases
7-Length of time to be indexed	

Splitting Down the Text. Splitting down the text as far as it makes sense and does not harm the integrity of the whole idea, may help to improve indexing to reach better retrieval. Clustering the text, says Hynek, [24] will maximize intra-class similarity while minimizing interclass similarity. Domain Analysis (DA) may be a good methodology for getting better reliable knowledge from the text. DA, as Roseti and Werner [25] denote, deals with knowledge identification, elicitation, and representation of software products, techniques that systemize these tasks are necessary.

Classifier/Indexer as Co-Partner of Reference Librarian. To reach this goal, although some efforts exist in some special libraries and information systems, I think it must be promoted at such a level that instead of training traditional technicians in the field, they can be regarded as a researchers and investigators. Also, the classifier/indexer must not be behind the scenes in libraries and information centers, but should have real contact with the users and researchers. As for reference librarians, one of the most applicable techniques is the reference interview. Here, for the classifier/indexer, there should be some possibility of

understanding the user's need and background. As a proposal, it may suffice if, when opening the interfaces, before any action, some applicable questions regarding the specialty, interest, level of the knowledge wanted, priority, and so on be asked.

Developing More Finding Aids. Finding aids are very important in the retrieval process and they should be well defined and always updated. Finding aids comprise not only the hard copies in libraries and information centers, but also sophisticated help systems in computers. They also should be available in various forms, such as help screens with more graphics and images. Even in databases, the help system sometimes does not work properly. One should spend a lot of time changing the words and phrases to adapt to one's need in the computer. Sometimes, there is real disappointment reaching the help.

Making a High-Level Thesaurus. Although a thesaurus may be treated as one of the finding aids, because of its importance it should be emphasized. A well-defined thesaurus, denotes Hynek,[26] can improve system's response significantly. Making a good thesaurus originating from the classification system and in accordance with subject headings, prevent the use of scattered terms among the users and researchers. That is, if all terms used for classification/indexing are consistent with whatthe thesaurus or subject headings suggest, it may be of great help to the researcher.

Clearing Up the Semantic Domain of Words and Phrases. Taking into view the ambiguity of index terms across cultures, languages, and time, it is evident that with concrete names there is no major problem in indexing and retrieval. What is more problematic is the indexing of abstract terms. If we can clarify multi-meaning words with some sort of explanations, there may be more hope to retrieve them more easily. It is true that there are many multi-dimensional words, and that the people use words and expressions with no concern for specialists in their application. But what we can do is to minimize the misunderstanding of the terms. This may happen, especially in technical domains, by specialization in the fields we are working on.

Standardization of the Terms Used. Although people use language and nobody can dictate to them how to use it, scientific terms are rather different from popular words. So, in scientific texts, if we try not to scatter keyword terms and bring them to some sort of standard level, it may work. In my article *New Scheme for Library Classification*,[27] I illustrated that if we harmonize the terms we use in our classification of knowledge–library classification terms, subject headings, thesaurus and indexing terms–it certainly results in conformity and stability, which would help greatly in retrieval.

Weighting the Terms. It is possible to make some feature for words. I have shown in the above-mentioned article, that every term in the indexing system can have two digits as its feature. These features help the user in retrieval. Referring to Losee's *Gray Code [28]* tabular form, we can make indexes with designated features as well. These numbers can be treated mathematically as the dimensions of the term. In this manner, every term has its own value, and can be defined by it. This is helpful for the researcher in choosing what s/he wants, and it prevents one from wandering around aimlessly. Suppose one wants to search under classification and this term is used in several disciplines such as Philosophy, Biology, Zoology, and Library Science. In searching, if one clicks on the term, it may appear on the screen:

Philosophy	Biology	Zoology	Library Science
5,6	8,1	10,1	13,5

One can then choose the desired discipline.

External

Promoting the Ability of the Searcher. Although it is true that in information retrieval everything is for the sake of the user and his consent is the ultimate goal of all classification and indexing systems, to reach this goal, the user should know that one has to prepare for a search. Searching is very easy if one knows its rules and it will be frustrating if one is not well acquainted with them. The Internet, says Hynek,[29] is a convenient and less consuming way to conduct research if [and only if] the user has the necessary literacy and equipment. So, to reach this goal, every library, as well as databases, must prepare some useful and userfriendly teachings for the users in order not to allow them to be frustrated, going here and there searching for their needed information.

I, in my personal experience, searched many times for the terms classification of knowledge, organization of knowledge, the philosophy of classification/organization of knowledge and also similar combinations for library studies in several engines as Google, Yahoo, All the Web, and Alta-Vista. The result was too many answers with high tolerance among them. Certainly, overlap is also a major problem. With a simple look at Table 2, one can imagine how difficult a task it is to search through Web sites for these terms. Some of the problems may be as follows:

- *The differences between search engines are amazing*. In this example Yahoo is the least container of all. The reason may be that it is not well equipped with scientific topics in comparison with other engines, especially Google.
- *Huge amount*. Searching millions of databases, of which most are leading to other databases, is extremely time consuming and frustrating. Meanwhile there is a lot of repetition in one engine, rather than other ones. Grossman and Frieder[30] reciting from [Kahle, 1998, Lawrence and Giles, 1998], say:

  > It is estimated that the Web now contains more than twenty million different content areas-presented on more than 320 million web pages, and one million web-servers, and it is doubling every nine months.

- *Selection by chance*. Searching all of the databases in one engine rather than all other engines is frustrating. One usually may choose some of the first entries by chance and one may not be sure to have reached all possible existing knowledge.

TABLE 2. Comparison of Items Retrieved in Several Engines

Engine / Subject	Google #	Yahoo #	All the Web #	Alta-Vista #
Classification + Knowledge	1,460,000	1	1,967,179	655,164
Classification + Knowledge + Philosophy	248,000	211,000	463,336	80,321
Organization + Knowledge	3,260,000	401	9,921,796	3,039,415
Organization + Knowledge + Philosophy	768,000	8	1,819,407	444,522
Library + Classification	1,090,000	3 + 18	1,556,345	429,207
Library + Classification + Philosophy	157,000	132,000	327,120	398,352
Library + Organization	2,520,000	20	8,306,760	2,484,141
Library + Organization + Philosophy	527,000	1	1,256,523	398,352

Seeking Real Human Information Needs. One must not forget that although highly-automated technology dominate, all things are for human beings. In fact, the profits of those who are behind the market and reap benefit of technology, although not admittedly, in practice have forgotten man and his real needs. Therefore, every information provider may seek his own superiority and economic benefit and may want to omit his rivals. Each one tries to attract clients all around the world, as many as possible. Although, competition amongst the rivals results in a good deal of innovation and scientific progress, what is the value of this innovation and progress in comparison with human losses? It is certainly a critical issue. The question is, if the real needs of people were the producers' aim and objective, could they not have innovations in another direction too?

More Human Intermediaries. It is true that information technology (IT), has attained its highest position these days and has served us well for searching and retrieval of information needs. But, there is still a high level of necessity for human intermediaries. In other words, the defect of highly sophisticated information systems is the negligence of human resources as providers, decision makers, users, communicators, consultants and so on. Thus, information on how to use and how to apply is not enough in comparison with information production. That is, information systems or services (IS), are not regarded by the same standard. So, it is common that the communication between the classifier/indexer and user is not at an acceptable level. One of the reasons that the user is not well satisfied is because of a lack of good communication, or the difference between his need and what the classifier or indexer has anticipated. Defining a job vacancy for information retrieval as retrieval consultancy is highly recommended.

Ranking the Level of the Document. As there are too many information providers, it is not clear to which discipline each relates and who their exact audiences are. In other words, as there are many sites containing too much information, there is no evidence as to what degree each of the site's information has actual relevance. So, some methods should be found to possibly distinguish the level of their subject relevance. It is said that Google, as an information engine, prioritizes the documents. That is, when you click on an item, it automatically brings up the most important ones in the 10 or 20 items shown first. Although I have not any proof for this, I think it is possible to find a way in which during manual indexing or in automatic searching, one manages to receive the most relevant items first.

More Cooperation from Information Providers. If information providers, from those who make databases to those who provide search engines, attempt to define their realm of expertise and not to duplicate other sites by interfering in their subject domains, it would be a great improvement. For example, separating commercial language from scientific or cultural language would attract related audiences and work as special libraries do. In addition, providing cross-references from one engine to another in the opening interface would benefit the user. In fact, this should be done after the full orientation of every engine to what it is responsible for. Thus, by the time you open the engine or even the site, you are well aware of exactly what the engine or site comprises.

Define or Create Some Sort of Software to Anticipate the User's Need. This may be done through enforcing the feedback system. A study of behavioral science and how the user makes decisions is of high importance, and full of difficulties. All of the researchers are fully occupied with investigating how to define this issue. Fenstermacher and Ginsburg,[31] mentioning Data Mining project explain:

> By expanding client-side analysis to monitor applications outside the browser, data mining can reveal patterns of behavior around Web access, and not simply within it. This field has been very active in both academia and industry.

However, I think trying to filter noises in communication and separate the commercial propaganda from the real, sophisticated domain of research may greatly help to anticipate the user's needs. The implementation of an easy quick feedback system, in whatever way possible, may facilitate the case.

CONCLUSION

Although a lot of work has been done, and information providers, as well as librarians, try to satisfy users, a lot of work remains. Taking into consideration all above-mentioned recommendations, there may be some other obstacles producing problems for better communication. The viruses, for example, show that there are some people who have devoted their lives to making problems and depriving others from accessing their information needs. In such a complicated condition, classifiers/ indexers try to interpret or estimate the users need, but how this can be done. As the elimination of viruses and virus makers is impossible, much effort must be devoted to filter out unwanted information. If classifiers/indexers communicate with the users in any means possible and receive their interest, they will certainly be able to help them. Communication in its best way possible is the highest task of human relationships, which can fill the gaps and accelerate the circulation of information fluently. This means that although information helps people to understand each other in a more appropriate way, the effective communication needs some loyalty and fidelity towards real human needs.

NOTES

1. Jennifer Rowley and John Farrow. (2000). *Organizing Knowledge* 3rd ed. London: Gower, p. 101.
2. Jiri Hynek. (2002). *Document Classification in a Digital Library*. Technical Report no. DCSE/TR 2002-04, 40 p., p. 20. Available at: http://www.kiv.zcu.cz/publications/2002/tr-2002-04.pdf
3. David A. Grossman and Ophir Frieder. (1998). *Information Retrieval; Algorithms and Heuristic*. Boston: Kluwer Academic Publishers, p. 5.
4. Don Lathom. (2002). "Information Architect: Notes toward a New Curriculum," *JASIST* 53(10):827.
5. Ricardo Baeza Yates and Berthier Ribeiro-Neto. (1999). *Modern Information Retrieval*. New York, London: ACM Press; Addison-Wesley, p. 19-21.
6. Grossman and Frieder, *Information Retrieval*, p. 2.
7. Grossman and Frieder, *Information Retrieval*, p. 11.
8. Robert M. Losee. (1990). *The Science of Information: Measurement and Applications*. New York: Academic Press, p. 227.
9. Losee, *Science of Information*, p. 195.
10. Masse Bloomfield. (2001). "Indexing–Neglected and Poorly Understood," *Cataloging& Classification Quarterly* 33(1), 66.
11. Losee, *Science of Information*, p. 195.
12. Losee, *Science of Information*, p. 197.
13. Rowley and Farrow, *Organizing Knowledge*, p. 134; Baeza and Riberto, *Modern Information Retrieval*, p. 27.
14. Hynek, *Document Classification*, p. 7.
15. Hynek, *Document Classification*, p. 8.
16. Ronald R. Yager. (2000). "A Hierachal Document Retrieval Language," *Informational Retrieval* 3:357-358.
17. Losee, *Science of Information*, p. 222.

18. Elaine Svenonius. (1992). "Classification: Prospects, Problems, and Possibilities," in *Classification Research for Knowledge Representation and Organization*, p. 5-28.
19. Rowley and Farrow, *Organizing Knowledge*, p. 24.
20. Baeza and Riberto, *Modern Information Retrieval*, p. 6.
21. Bloomfield, "Indexing," p. 67-70.
22. Hynek, *Document Classification*, p. 18.
23. Hynek, *Document Classification*, p. 20-21.
24. Hynek, *Document Classification*, p. 30.
25. Monica Zoperlari Roseti and Claudia Mara Lima Werner, "A Knowledge Acquisition Systematic Within the Domain Analysis Context", p. 1. Available at: http://www.cos.ufrj.bd~odyssey/publicacoes/wer99f.pdf
26. Hynek, "Document Classification," p. 9.
27. Fadaie Araghi, Ghoamreza, (2004). "New Scheme for Library Classification," *Cataloging & Classification Quarterly*, 38(2).
28. Robert M. Losee. (2002). "Optimal User-Centered Knowledge Organization and Classification Systems: Using Non-Reflected Gray Codes," *Journal of Digital Information* 2(3), March. Available at: http://jodi.ecs.soton.ack.uk/Articles/v04/i01/Losee/
29. Hynek, "Document Classification," p. 34.
30. Grossman and Frieder, *Information Retrieval*, p. ix.
31. Kurt D. Fenstermacher and Mark Ginsburg. (2003). "Client–Side Monitoring for Web Mining," *Journal of the American Society for Information Science*, 54(7).