Major Problems in Retrieval Systems

Gholamreza Fadaie Araghi
University of Tehran, Department of Library and Information Science

ABSTRACT

For better retrieval, classification and indexing are the key factors. To better serve users there should be some criteria. The most popular criteria known are recall and precision. But these two are not totally accepted and respected in application. Uncertainty and giving more respect to information technology at the expense of information system management are the main problems. Although information retrieval includes many elements such as modality, document classification and categorization, system architecture, user interfaces, data visualization, filtering, language, and behavioral sciences, communication in a good environment covers all of them. For better communication, a classifier/ indexer must avoid taking false direction, be frank and careful not to use ambiguous terms, and must provide sufficient cross-references in their indexing. Good questioning and saying no when it should be said is regarded as a kind of filtration of the environment from noises.

KEYWORDS: Recall, precision, filtration, uncertainty, communication, information system

INTRODUCTION

Although the field of information retrieval has made much progress, many problems still exist. Those who provide information or manage it must take these problems into full consideration. Indexing and classification are the most commonly used tools to answer the user's need. Some advanced systems for better retrieval such as Boolean, Vector, and Fuzzy approaches are developed to cope with the problems. But there is still doubt that these approaches and systems can highly promote the efficiency of the task. To evaluate the retrieval process, recall and precision are the most popular methods known at the present time. But some think that they do not work properly. While uncertainty is a major obstacle on the way to answer the user's need, the efforts of information providers are devoted mostly to the process of information technology (IT). Although Information technology is of high importance, it must be used totally to serve needs. Information system (IS) management not only should be regarded in the same way as information technology but we must assign it some priority. That is, if we allocate some money and energy for IT, we must allocate more for IS. It is critical to serve users with least investment in IT in order to get more benefit in information system management. The economics of information, which is getting more attention these days, recommends this. The evaluation of a system depends on the extent of services we get from the amount of investment we allocate for information technology.

RECALL AND PRECISION

Although there are various methods to evaluate the retrieval process as well as classification activity [1], recall and precision are highly recommended by the authors. The disputing opinions on this range from recall and precision being nonsense and completely rejected to nearly full acceptance.[2] Regardless, as mentioned before, the

satisfaction of the user in the retrieval process is to be shown by relevance. And recall and precision are highly connected to relevance and non-relevance. Bloomfield [3] argues that there is no advantage to using recall and precision. One of the major reasons for the inapplicability of recall, he says, is because we do not know the exact number of relevant items in the whole database. So recall which means relevant items retrieved in relation to the whole number of relevant items in the system, actually becomes impossible to calculate and unreachable. Precision, too, is defined as relevant items found in relation to the relevant items found plus the irrelevant items found by the user. Bloomfield argues that non-relevant items found by the system are not really counted as retrieval. Retrieval, practically, means those relevant items, for which the user is looking. If the system retrieves some items that are not relevant, it is a defect of the system and a wasting of time for the user, and it is not effective retrieval. So here we do not see any advantage for precision except that it is equal to retrieval itself. As Maltby says:

> Recall depends on many factors including depth and accuracy of indexing, but attempts to achieve greater precision involve the use of controls of various kinds and these often are distinctly classificatory in character.[4]

And Rowley and Farrow [5] state that recall depends on the system's ability to filter out unwanted items. They mention [6] that these two are capable of being measured under controlled conditions, and they are used to express them by ratios. They count hits, *miss*, *noise*, and *dodge* for the system; in a good system one should minimize the *noise* and *miss* in order to get more hits. The indexing system and search software, they emphasize, are the means to maximize recall and precision.

What is known from the statements of the above mentioned authors who still believe in recall and precision may be categorized as follows:

- Recall and precision are a traditional measure for retrieval qualification.
- They are one of the evaluation measures and may be the simplest one.
- There is a classificatory measure in them. This means that if we maximize the potential of the classification/indexing system we can be more hopeful of fulfilling our needs.
- They are ideally measured under controlled conditions.
- They are usually inversely related to each other. That is, by broadening the search we have improved the recall but at the cost of lower precision.
- By minimizing *noise* (retrieved irrelevant items) and *miss* (relevant items not being retrieved) we can maximize both recall and precision.

But the question is: What are recall and precision and is it possible to maximize both? Another question is: Are recall and precision not the same as specialty and generality? The answer to the first question may be that recall and precision are nothing but retrieval itself. Because, if the classification/indexing system does its job well and the user is well acquainted with the system, all of the retrieved items are those which are relevant and wanted by the user.

What are called *noise* and *miss* actually are not supposed to be called wanted information retrieved. They are just like Spam, which everybody tries to get rid of. They are nonsense or waste information, and may be considered a defect of the system. In a perfect system, as pointed out by Grossman and Frieder, [7] only relevant documents are retrieved. This means that at any level of recall, precision would be 1.0. Recall and

precision may be maximized by an increase in speed and easy access. In these two ways, that is, in the case of a reduction of *noise* and *miss*, we can ensure the maximization of the system's quality. What one expects from the system is to be able to carry out one's query in as short a time as possible. In other words, relevant items in a limited time, which implies easy access by using the best classifying/indexing system through users instructions, provides a measure for evaluation. Believing in recall and precision, the best system is the system in which recall and precision are both high. Suppose that you have five documents that you have classified or indexed in the best way possible. If the searcher is well acquainted with the system, all of the five items should be found and recall and precision are the same. But when the number of documents increases and there is a gap between the user needs and what the classifier/indexer does, because of lack of communication, the ratio of precision and recall differs and may decrease.

The answer to another question may be that if the system works well, a knowledgeable user with well-defined needs should retrieve every relevant item. So if one does not find his/her needed information it may imply that the particular database does not include that specific subject. One may have to move to a broader subject to find more general topics. This means moving from a more specific to a more general subject and does not relate to recall and precision. In fact, this may be called the generality and specialty referred to earlier. Defining A. G. Brown's idea, Taylor notes that [8] if we go in-depth with indexing we may get the subtle information, but if we summarize, we can only reach the general concepts in the first step of the retrieval process.

Finally, if we still want to use the terms recall and precision we must know that, in a good system, they are not opposite to each other and they can both attain an optimum point, which is the best for the user. They can be in the same direction, too. So, what causes to differentiate these two as a measure of quality control may lie in the indexing/searching system.

THE EXISTING PROBLEMS

Coming back to the indexing system, and usually with the lack of information about the user's needs and behavior, some major problems exist with methods for retrieval. The main ones are:

Uncertainty. The consent of the user is the major problem. Going through user's consent leads us to the study of his search behavior. And this, in turn, leads us to the uncertainty and probability of one's decision-making. Stating user's behavior and understanding his information needs highly affects the organization and operation of the information retrieval system [9] and may help us in predicting his decision-making. Decision-making is connected to many factors in practice. Decision science as defined by Brugha [10] is mostly related to philosophy, information system, psychology, culture, and management. And this is the reason that prediction of what the user wants or decides becomes very difficult. Brugha shows that the user's need and what he is looking for relates to his type of thinking.[11] Taylor explains that the subject approach has become a major way of finding information in the electronic age.[12] Search engines have tried to fill the void on the Internet, yet users became more frustrated with the thousands of so-called "hits" beyond their desired results.

As the classifier/indexer is not in the same environment where the user may be, although one may try his best in this domain, the differences in the time and place may affect his work. We must accept that both classifier/indexer and user are decision-makers in their job and in the special environment they are in. For example, the cultural and scientific difference between classifier/indexer and user may economically affect the task. And from this angle, a new field of study has appeared, called the Economics of Information.

Information System vs. Information Technology. The superiority of Information technology (IT) over Information systems (IS) roots from the same logic that the introductory means dominate the ultimate goals. As was discussed before, classification may determine retrieval in the same way that production precedes need in Economics. The priority of classification and retrieval, as well as production and consumption, goes back to the debate of the priority of want and need. Because of too much production by some special groups of people, others must apply them. Marketing is ultimately a job which entails more clients for this production. Flaming up the wants, but not needs, generates the motive for more and new production. Therefore, the idea for too much production becomes superior to good consumption. For this reason, every project for IT is to satisfy humans' ambitious desire, while fewer attempts are made for IS to use them in a natural way according to general human needs. Meanwhile, the information acquired in this way is sent to users to stimulate their appetites and desires. This sending of everything to everybody may cause information traffic or "information pollution." Too much unwanted information really makes for frustration. If there is an equation for production and consumption on the basis of human needs, there may be less frustration. Spam is an example. So, trying to send the right information to the right person at the right time, as is the exact meaning of information science would certainly take more time and be more expensive. We may all have experienced searching for things that were never found or were found after the deadline.

Pinto and Millet [13] say that whenever greater budgets are provided for IT, less advantage come in IS. Information systems in our organizations have arrived at an unbelievable condition. Some statistics mentioned by them confirm this:

- A recent study of over 300 large companies shows that software or hardware developments fail at a rate of 65%.
- Half of IT projects become runaways while failing to deliver fully on their goals.
- Up to 75% of software projects are cancelled.
- Of approximately 17,500 projects costing more than $250 billion each year, 52.7% will overrun their initial cost estimates by 189%. Most of these projects are delivered with only 74% of original functionality.
- A U.S. Army study of IT projects found that 47% were delivered but not used; 29% were paid for but not delivered; 19% were abandoned or reworked; 3% were used with minor changes; and only 2% were used as delivered.

Bloomfield [14] provides some more evidence about using the Internet and argues that searching devices can be frustrating. By bringing some evidence from various journals emphasizing better indexing, he notes that there is a lack of a theoretical foundation in our way of developing indexing procedures.

MORE ABOUT COMMUNICATION

Information retrieval includes many elements, says Baeza Yates and Ribeiro- Neto, [15] such as modality, document classification and categorization, system architecture, user interfaces, data visualization, filtering, language and behavioral sciences, communication and others. But, I think that in good communication, all of the elements are involved. The main point to focus on, insofar as it is more applicable and attainable in a short or medium time period, is indexing.

In order to communicate better with the user and to avoid taking false direction, the classifier/indexer must be very frank and careful not to use ambiguous terms, and must use sufficient cross-references in his/her indexing.

Before concluding, I would like to look at two Persian-Arabic proverbs, which say:

1. La adri nisf al ‘ilm (to say I do not know when you really do not know is equal to half knowledge) [16]
2. Husn al su’wal nisf al ‘ilm (good questioning may be regarded as equal to half of the knowledge) [17]

These two proverbs relate to retrieval rather than the classification/indexing procedure. Emphasizing Su’wal (the question) brings out the importance and priority of the question in comparison to information, and this is what Lauer18 states about Churchman’s inquiring systems as a question-centric approach to knowledge management. Further explanation of these two sentences is as follows:

(a) a question which originates from everybody’s need is the main core in our educational life,

(b) one should enquire from someone or some reference source to which the answer belongs,

(c) if the source referred to–human, library book or database–does not have the answer, it must frankly admit this and say no,

(d) this would benefit the source itself. That is, if the source, human being, text or database, needs the information, it must be provided, and

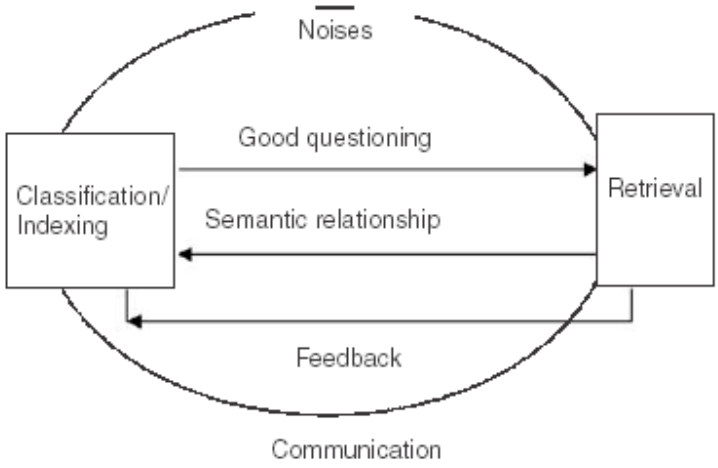
(e) The user saves time by searching the new relevant source for the answer instead going in the wrong direction. Saying NO when it should be said is regarded as a kind of filtration of the environment from noises. These noises are the main enemy of communication. Nowadays, in order to attract more clients and make more profits, marketing by using all persuasive phrases and words usually goes far beyond what they actually represent. This, in fact, is misleading instead of leading, and misunderstanding instead of communication.

Someone may ironically say that with these two proverbs, the seeker may attain the whole of knowledge because one has captured two halves of the knowledge. The answer is that developing a good question, searching to find the right resource, plus a viable methodology to get the result is quite enough to get nearer to the answer. That is, one may find the answer sooner than one expects. In other words, taking these two pivots into consideration, the right communication is possible. If everybody emphasizes questioning, then tries to find the appropriate resource, human, database or other, one will soon find the answers. Lauer [19] defines very well how asking a good question affects finding the relevant answer. Discussing Churchman’s idea in his article on information or question preference, Lauer [20] argues that our life activities must be question-centric rather than

information-centric. Otherwise, one may find a lot of information which is not required, and therefore, one soon loses or forgets it. In these days, everybody is frustrated with the huge amount of Spam. Spam, by itself is information, but as it is not fulfilling one's personal, social, or special need, one has to spend a good deal of time getting rid of it. In a rich learning society, Lauer reveals [21] that knowledge flourishes, and a society based on questioning will be more progressive than an information-centric one.

The retrieval model based on the question-answer process is a kind of communication. In a communication system, if the sender and the receiver know each other as much as possible and try to diminish the parasites, they will get better results. Figure 1 illustrates this.

FIGURE 1. The Relationship between the Classifier/Indexer and Retrieval



Taking Figure 1 into consideration, we can imagine the importance of communication and the human relationship in information retrieval. In fact, there is much similarity between the environments in which the classifier/indexer and user work and the model of the communication procedure as Shannon and Weaver proposed.[22] Here, the classifier/indexer acts as sender and the user acts as receiver, and the library or database is the same as the conveyer or transmitter. In this case, we should minimize the parasites and noise as much as we can. The clarification of language, semantically and grammatically, for both sides (indexer-user) is of high importance. Experiencing all indexing techniques to promote the best results is recommended. For example, Green [23] proposed her frame-based language index system to get better results in the expression of systematic relationships. Latham [24] has defined communication as a major task in the new curriculum for information architecture. He says that the function of information architecture is the same as that for information retrieval. For good communication, one should be acquainted with the psychological, social, cultural, economic, and religious aspects of human beings.

It is also necessary to use some techniques used in reference services, such as the reference interview, to welcome the user to explain his information need. In communication, especially in the age of the information explosion, there is a huge amount of noise. Everybody adds his own words and tries to attract more clients by every means possible. One uses any metaphor and he may create some new ones to kick other

rivals out of the competition. Although this may be considered a positive procedure to them, as everybody finds some share in the global communication system, it is harmful in decision-making.

This rush to convey one's message results in a sort of anarchy and creates a lot of trouble for the supposed special audience trying to retrieve relevant messages easily. After globalization of information and thinking of the world as an information village, I think now is the time for partitioning global knowledge into related sections for the use of related users as a turning point. In this dividing process, more action on the side of classification/indexing may be necessary.

CONCLUSION

Retrieval depends primarily on classifying/indexing. But the main thing is the way we look at it. If our view is dynamic, that is, if we classify/index to retrieve, then everything may change. And if, in theory or in practice, we classify/ index because it is a job and we are told to do so, nothing will change.

Even in the case of information engine providers, although they try to satisfy users by gathering databases, it seems that their main idea is to attract audiences' attention by their abundance of information, not by methodology and their help systems. This is the same as a static view to the library and information system, in which the accumulation of information is more important than successful retrieval. It means that every library and information system, as well as information databases, in order to become a super power as an information provider, tries to increase its assets by collecting that which relates or does not relate to it. This may also be because there has never been a clear definition for their activities. In such situations, serving the clients may be considered to be the secondary task. Another factor may lie in the fact that the libraries, databases and information providers are not established primarily for the sake of needs, but rather for the sake of wants. And although wants lead to more creation, bring research and development, and initiate new activities, they may be beyond actual service to the real needs of clients. That is, certain companies for the sake of fulfilling their ambitious projects perform activities which are potentially harmful. Afterwards, others must use the results. This strategy actually has resulted in the superiority of IT rather than IS. In other words, some special companies create information; others have to use whatever is provided for them. But here, too, they are rivals with each other and everybody should choose one's way heuristically.

Following the real aims and objectives in information retrieval, besides having a clear strategy based on human needs for information, classification and indexing is of high importance. Human resources in the form of retrieval consultants, in order to facilitate the process of developing a better retrieval system, are recommended.

REFERENCES

1. Jiri Hynek. (2002). *Document Classification in a Digital Library*. Technical Report no. DCSE/TR 2002-04, p. 20. http://www.kiv.zcu.cz/publications/2002/tr-2002-04.pdf.
2. Arthur Maltby. (1975). *Sayer's Manual of Classification for Librarians.* 5th ed. Gt. Brit.: Andre Deutsch/Agrafton Book, p. 309; Jennifer Rowley and John Farrow. (2000). *Organizing Knowledge*, 3rd ed. London: Gower, p. 341; Masse Bloomfield.

(2001). "Indexing–Neglected and Poorly Understood, *Cataloging & Classification Quarterly*, 33(1):69; Hynek, *Document Classification*, p. 5; David A. Grossman and Ophir Frieder. (1998). *Information Retrieval: Algorithms and Heuristic*, Boston: Kluwer Academic Publishers, p. 1-9.

3. Masse Bloomfield. (2001). "Indexing–Neglected and Poorly Understood," *Cataloging & Classification Quarterly*: 33(1):70.
4. Maltby, *Sayer's Manual*, p. 309.
5. Rowley and Farrow, *Organizing Knowledge*, p. 341.
6. Rowley and Farrow, *Organizing Knowledge*, p. 362-3.
7. David A. Grossman and Ophir Frieder. (1998). *Information Retrieval: Algorithms and Heuristic*. Boston: Kluwer Academic Publishers, p. 4-5.
8. Arlene Taylor. (1999). *The Organization of Information*. Inglewood, Colorado: Libraries Unlimited, p. 135.
9. Ricardo Baeza Yates and Berthier Ribeiro-Neto. (1999). *Modern Information Retrieval*. New York; London: ACM Press, Addison-Wesley, p. 7.
10. Cathal M. Brugha. (2001). "Implication from Decision Science for the Systems Development Life Cycle in Information Systems. *Information System Frontier*, 3(1):92.
11. Brugha, "Implication from Decision Science," p. 93.
12. Taylor, *Organization of Information*, p. 131.
13. Jeffrey K. Pinto and Ido Millet. (1999). "Successful Information System Implementation, the Human Side," 2nd Edition. Project Management Institute, 1999.
14. Bloomfield, "Indexing–Neglected," p. 72-3.
15. Baeza Yates and Ribeiro-Neto, *Modern Information Retrieval*, p. 7.
16. 'Ali Akbar Dehkhoda. (1372/1993). *Amthal va Hekam (Axioms and Parables).* Tehran, Iran: Amirkabir, pt. 3, p. 1340.
17. Dehkhoda, *Amthal va Hekam*, pt. 2, p. 994.
18. Thomas W. Lauer. (2001). "Question and Information: Contrasting Metaphors," *Information System Frontiers* 3(1):41-8.
19. Lauer, "Question and Information," p. 47.
20. Lauer, "Question and Information," p. 47.
21. Lauer, "Question and Information," p. 47.
22. Warren Weaver and Claude E. Shannon, *The Mathematical Theory of Communication*, Urbana, IL: University of Illinois Press, 1949, republished in paperback 1963.
23. Rebecca Green. (1991). "The Expression of Syntagmatic Relationships in Indexing: Is a Framework Based Index languages the Answer?" in *Classification Research for Knowledge Representation and Organization*, edited by Nancy J. Williamson and M. Hudon. New York: F.I.D.; Elsevier, p. 79-89.
24. Don Lathom. (2002). "Information Architect: Notes Toward a New Curriculum," *Journal of American Society of Information Science and Technology*, 53(10):825-828.