What do you say after you say, 'I work in $\mathcal{IR}$' ?

Vijay V. Raghavan
The Center for Advanced Computer Studies
University of Southwestern Louisiana
Lafayette, LA 70504

1 Introduction

The discipline of Information Retrieval($\mathcal{IR}$) faces several important problems. One of these problems is that it lacks a clear identity. This problem has led to the following undesirable consequences. On the one hand, both people in the field and those outside have only a vague understanding of what $\mathcal{IR}$ really is. On the other hand, since the introduction of new terminology has not kept up with the actual progress made in terms of new concepts and methods, there is an impression that the field is stagnant.

Through a review and analysis of the various definitions of $\mathcal{IR}$ that exist in the literature, it is shown that the problem is real. It is argued that it is a problem that should be resolved in order to ensure that we project a good image and communicate better with the community at large.

2 A Medley of Viewpoints

Modern computer literature is abound with definitions of $\mathcal{IR}$ and other related terms. For the purposes of our discussions, a number of definitions are selected and grouped as follows: *Early work, From the insiders* and *From the outsiders*. Let us consider each in a separate subsection.

2.1 Early work

The definitions in this subsection are selected from [Koc74]. It is thought that C. Mooers was the first to coin the phrase 'information retrieval'. He states,

> $\mathcal{IR}$ system [is viewed][1] as a machine that indexes and selects information in a library.
>
> **C. Mooers (1951)**

A somewhat more accurate description is given in the following:

> $\mathcal{IR}$ system [is] a way of providing people with documents they need.
>
> **H. P. Luhn (1952)**

The importance of the distinction between information-retrieval and literature-search is stressed in the next definition.

> Literature-search produces a list of references in response to a topic description. This is also called *document retrieval* by some.
>
> $\cdots$ Literature-search retrieves references to units likely to contain the sentences that information-retrieval would display directly.
>
> **Bar-Hillel (1960)**

Kochen provides further elaboration of these ideas in the following definition.

> [Information Retrieval involves the] production of a list of declarative sentences in response to interrogative sentences. $\cdots$ The statements may be displayed directly as responses to expressed needs, or they [the stored statements] may steer a computer to generate responses to requests [Q-A systems]. $\cdots$ We will use fact-retrieval to mean answering questions that do not require inference.
>
> **Kochen (1969)**

2.2 From the insiders

In the following definitions, from [vRij79, Mea73, Hea78], one discerns a clear shift in the meaning of information retrieval from those given by Mooers, Bar-Hillel or Kochen.

[1] Square brackets in quotations are added by this author for clarity.

> The function of a retrieval system ··· is to locate and recover information that is stored away. Often ··· the patron uses [a] ··· query to locate and recover some string of symbols which might ··· range from a single number to a shelf of books.
>
> **C. T. Meadow (1967)**

> In fact, in many cases, one can adequately describe ··· [information] retrieval by simply substituting 'document' for 'information'.
>
> ··· An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of inquiry. It merely informs the existence (or non-existence) and whereabouts of documents relating to his request. This, specifically, excludes Q-A systems. ··· It also excludes data retrieval systems ···· .
>
> **C. J. van Rijsbergen (1975)**

> Although ··· information usually is retrieved by means of stored data that represents documents, it is the emphasis on information relevant to a request, rather than direct specification of a document, that characterizes the modern subject of information retrieval.
>
> ··· The problem of locating relevant information from a body of widely dispersed knowledge is analogous to detection of the presence of a signal pulse in the presence of a noise background.
>
> ··· [there is] a requirement for more precise mathematical formulation of the principles of $\mathcal{IR}$ in order to ensure that computers and computer accessible storage devices are used in an economic manner. ··· The measure of relevance may be formulated in mathematical terms and leads to considerations based on the mathematical theory of pattern recognition.
>
> **H. S. Heaps (1978)**

2.3 From outsiders

The following definitions from text books on Database Management are noteworthy. They [Mar75, Dat83] provide a somewhat different perspective on what information retrieval is about.

> An important category of inverted file system is the document searching or text-retrieval system.
>
> $\cdots$ A search criterion may be [to] search for all documents **containing** both of two specified words, for example, $\cdots$ "COMPUTER" and "CHESS".
>
> **J. Martin (1975)**

> $\cdots$ [In] text search or information retrieval applications, in which the database contains (for example) scientific abstracts or other textual information, the overall structure is much less regular. Queries against this kind of database tend to be quite complex. $\cdots$ Such a query requires the system to scan long text strings, looking for occurrences of substrings such as "DATABASE MACHINE" or "ASSOCIATIVE DISK" or "CELLULAR LOGIC".
>
> **C. J. Date (1983)**

3 Discussion of Problems

The definitions of the previous section give us some idea of the seriousness of the problem. In the following subsections, two problems resulting from these definitions are discussed. The first section addresses the ways in which the definitions are inconsistent. Then, some directions for the nature of changes that might be considered are presented.

3.1 Lack of Clarity

The existing viewpoints disagree in several respects. First there is the question of how $\mathcal{IR}$ relates to fact retrieval, reference retrieval and question-answering. For example, Bar-Hillel and Kochen exclude reference retrieval from being part of information retrieval. In addition, they both consider fact retrieval to be a part of information retrieval. Kochen goes further to also include question-answering as a type of information retrieval. Meadow generalizes in another direction in that he deems reference retrieval, as well as fact retrieval (and, perhaps, Q-A), to be $\mathcal{IR}$. In contrast, Luhn, van Rijsbergen and Heaps(LRH) exclude both fact retrieval and Q-A from their definitions. As far as the output to be generated by the system is concerned, the views of Martin and Date agree with those of LRH (even though only in form).

Secondly, there is the question as to whether information retrieval is viewed as a technical area or a behavioral area. In order to consider this issue let us introduce the distinction made by Salton [Sal89] between technical and behavioral areas of information processing. The *technical* area deals with information representation and manipulation, including methods of introducing and storing information in computers and $\cdots$ making them accessible to interested users. The *behavioral* area is concerned with the accuracy associated with the retrieved information in conveying intended meanings and the effectiveness with which it affects users' conduct.

Clearly, the views of Mooers, Meadow, Martin and Date treat $\mathcal{IR}$ as a technical problem. Specifically, the problem is defined as one of searching for certain string patterns in a textual document. On the other hand, in the views of LRH, $\mathcal{IR}$ is a behavioral problem in the sense that what is retrieved (and thus deemed relevant by the system) may not necessarily be relevant (as determined by the user) and vice-versa.

Furthermore, one can observe that the view of $\mathcal{IR}$ differs not only between the *insiders* and the *outsiders* but also just among the *insiders*. Specifically, if we treat Kochen as an insider, there is really a wide spectrum of meanings covered by Kochen, Meadow and LRH. Although a majority of insiders may agree with the views of LRH, there are now some developments that cast some doubt on this position. For example, recent work by Salton does various kinds of local analysis to determine the relevant of portions of a document. This kind of work can clearly lead to a form of fact retrieval [SaBu90]. In addition, although the main focus of the retrieval task based on hypertext systems concerns the retrieval of chunks (of text), it may be possible to achieve fact retrieval by making the chunks sufficiently small. Thus, the question of what tasks should be included in $\mathcal{IR}$ is taking on a new twist.

3.2 New Terminology

Recently DARPA initiated a research project on *text retrieval and understanding*. The goals of the project certainly overlapped with those of typical $\mathcal{IR}$ research. However, it is rather curious that DARPA's RFP never made reference to the phrase 'information retrieval' [Dar90]. Instead, the requirements are stated as

> algorithms are desired for: (1) Detecting documents (and portions of documents) dealing with topics of interest. (2) Extracting specified data from documents.

Further clarification of (1) stated that

> applications for requirement (1) are automatic routing (for incoming documents) and retrieval (for retrospective searches).

It was, perhaps, felt either that it is less confusing to introduce new terminology or that the existing concepts are inadequate. Although the introduction of new terminology may be necessary, one must be cautious. After all, it is of no value to introduce new terminology just to be fashionable. Do we really need phrases like *message routing* and *retrospective searches*? What about *old* phrases that were generally popular and well understood such as *Selective Dissemination of Information* (SDI) and *On-demand searches*? In fact, there already exist a whole lot of buzzwords that no one carefully defines; for example, *Intelligent Information retrieval, User-oriented information retrieval, Adaptive information retrieval and Concept-based retrieval.*

Discussions we have had with prominent individuals in areas such as Pattern Recognition and Database Systems suggests that our field may be due for a name change. For example, Prof. Chandrasekaran of Ohio State University is of the opinion that $\mathcal{IR}$, as defined by LRH, is not really document or information *retrieval*, but rather document [relevance] *recognition*. John Mylopoulos of University of Toronto reacted to a similar discussion by comparing Knowledge-based systems (KBSs) to Database Systems (DBSs). Specifically, he views DBSs as a complex of symbolic structures and a *performance theory*, whereas KBSs consist of symbolic structures and *semantic theory*. The point is that the retrieval of text according to whether certain keywords of interest appear in the text would be the counterpart of DBSs, since that problem involves the management of symbols in textual documents in order to provide efficient access to them via appropriate performance theory. On the other hand, retrieval of relevant documents requires semantic theory that relates stored symbols to the content of actual documents and provides the connection between system design and users' needs.

4 Recommendations

We believe that $\mathcal{IR}$ faces a sort of identity crisis. Providing a satisfactory solution requires the cooperation of many interested parties. It is, therefore, recommended that a publication such as the ACM Special Interest on

Information Retrieval's newsletter (SIGIR Forum) be used as a vehicle to encourage the discussion of issues raised in this paper. In particular, this author believes that a view such as that expounded by Heaps [Hea78] should be refined and popularized.

References

[Dar90] DARPA. BAA-DARPA Research on Text Retrieval and Understanding. *Commerce Business Daily*, Friday, June 1, 1990.

[Dat83] C. J. Date. *An Introduction to Database Systems.* Volume II, Addison-Wesley Publishing Company, 1983.

[Hea78] H. S. Heaps. *Information Retrieval- Computational and Theoretical Aspects.* Academic Press, 1978.

[Mar75] James Martin. *Computer Data-base Organization.* Second Edition, Prentice-Hall Inc., 1975.

[Koc74] Manfred Kochen. *Principles of Information Retrieval.* Melville Publishing Company, 1974.

[Mea73] Charles T. Meadow. *The Analysis of Information Systems.* Second Edition, Melville Publishing Company, 1973.

[Sal89] Gerard Salton. *Automatic Text Processing.* Addison-Wesley Publishing Company, 1989.

[SaBu90] G. Salton and C. Buckley. Flexible Text Matching in Information Retrieval. Technical Report 90-1158, Dept. of Computer Science, Cornell University, Sept. 1990.

[vRij79] C. J. van Rijsbergen. *Information Retrieval.* Second Edition, Butterworths, 1979.