Salton Award Lecture
Information Retrieval as Engineering Science

Norbert Fuhr
Faculty of Engineering Sciences
University of Duisburg-Essen
Duisburg, Germany
norbert.fuhr@uni-due.de

Categories and Subject Descriptors

H.3.0 [**Information Storage and Retrieval**]: General

General Terms

Theory, Experimentation

Keywords

theoretic foundations, explanations, standardized testing

Extended Abstract

First, let me say how pleased I am to receive the Gerard Salton Award from SIGIR. His pioneering work strongly influenced our field for several decades. I had lively discussions with him at several occasions, especially during his stay with our research group in Darmstadt in 1988. We were following different theoretic approaches—he as inventor of the vector space model, myself as a young researcher being enthusiastic about probabilistic models—but the discussions with him improved my understanding of the commonalities and differences of both approaches, such as the role of representations, underlying assumptions and theoretic justifications.

Like my predecessors, here I want to give my personal view on information retrieval, by first presenting my own definition of the field, and then pointing out some areas for future research.

For me, IR deals with *vagueness and imprecision in information systems.* The imprecision is usually caused by the imperfection in the representation of the semantics and pragmatics of the objects stored, which are typically (multimedia) documents. Vagueness, on the other hand, is due to the fact that the user is not able to formulate her information need in a precise way; this leads to an iterative search process (which might even occur with databases for formatted data, like e. g. in product search in a Web shop).

This definition does not make the usual distinction between database (DB) and IR systems. In fact, in a certain aspect, I view IR as a kind of generalization of the former: The logical DB view interprets query processing as the task of finding those objects o for a query q for which the implication $o \rightarrow q$ holds. Rijsbergen defined IR as being based on uncertain inference, i.e. determining the probability $P(d \rightarrow q)$ that document d implies the query. Thus, from a querying point of view, classic DBs can be regarded as a special case of IR (though probabilistic DBs have become a major research topic in the DB field in recent years).

However, the major difference between DB and IR lies in the consideration of the pragmatic level: The major goal of an IR system is to support the user in solving her current problem or fulfilling her task at hand. Thus, standard IR evaluation is based on the concept of relevance or, in the case of user-based experimentation, on task-related criteria like task time or success rate. In contrast, the DB field delegates the pragmatic aspects to the application layer, which is beyond the DB scope.

Now let me turn to the major topic of this paper: IR as an engineering discipline. Certainly, many people working in this field would feel themselves as IR "engineers", i.e. they apply and extend known methods for solving new problems. However, a civil engineer planning a new building, a mechanical engineer constructing a new machine or an electrical engineer designing a new device—they all build upon a rich portfolio of basic findings and theoretic models, which not only allow them to solve the problem, but also make predictions about the result of their efforts (as well as knowing the limits of their methods). In analogy, assume an IR "engineer" being confronted with the task of designing a system for a new collection of documents and a new type of information needs—would she be able to guarantee a certain MAP or an average search task completion time? Rather not—instead, the standard practice is to try out some of the existing methods and tune them for the specific case—then an evaluation will show if and how well this system works.

So we might wonder what is required for constructing an IR system like in other engineering disciplines? I think we need two things that allow us to build upon, namely 1) theoretic models, and 2) solid empirical evidence.

On the theory side, the basis is rather small. E.g., for the case of ad-hoc retrieval, we have the probability ranking principle, various models based on it (like the classical probabilistic models, the principle of uncertain inference, and language models / divergence from randomness), and the term weighting axioms. These frameworks tell us *how* to achieve good or even optimum retrieval performance, provided that the underlying assumptions hold. However, we hardly have theories that tell us *why* certain methods work and others don't. From an engineering perspective, we would like to know how document features like length, language, domain, genre, structure or the term definition (single words, phrases, named entities etc.) affect quality

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

(and why)—and in a similar way, also for the topic characteristics as well as the relevance definition. Alas, we can hardly answer any of these questions.

On the empirical side, we have a vast amount of papers presenting numerous variants of existing models and methods, along with experimental results. This is partly a consequence of favoring experimental over theoretic work in our field. However, in the light of our general theme of IR as engineering science, this development was not helpful, for two major reasons:

1. We are not able to give a fully satisfying explanations of the outcomes of the experiments performed. Since we are mainly focusing on retrieval quality, too little attention is paid to the different factors affecting the results, like e. g. the representation used, the various assumptions underlying the retrieval model, the method used for estimating the model parameters, or the specific properties of the test collections used; often, good or bad performance is attributed to 'the model', without investigating its various components or the influence of the other factors (or their combination).

2. We can not predict to what extent these findings can be generalized to other settings: For a single collection, it is impossible to tell how far we can generalize from this single point of observation. Even with multiple collections, given that each one represents a point in a high-dimensional parameter space (and we even don't know what the crucial parameters are), we can hardly define the area where interpolation is reasonable—plus the problem that these are stochastic experiments anyway, and we would need more observations to get a statistically significant result.

In order to draw any useful conclusions from experimental results, we need standardized test environments and test procedures for being able to aggregate knowledge from larger numbers of experimental studies. I know only of a single meta-study [2] that looked at papers proposing specific methods for improving the quality of ad-hoc retrieval, which used some of the official TREC and CLEF collections. Although all of the considered papers claimed some performance improvements over previous work, hardly any of them was actually able to beat the best official run for the corresponding collection—the authors had been cheating by using poor baselines for comparison. This experience calls for more standardized experimentation, so that experimental results can be compared immediately.

Unfortunately, in recent years, we have seen a trend in the opposite direction: A large fraction of the papers at the top IR venues publish experimental results using proprietary collections. Thus, the experiments presented cannot be repeated by other researchers. This procedure does not satisfy the basic requirements for empirical scientific work, and it is kind of surprising how the IR community tolerates this violation of scientific standards. Proprietary data not only eases cheating (as we are no longer able to find out, this might even increase the frequency of this kind of misconduct), the major problem is the impediment of scientific progress: If someone has an idea on how to improve over the method presented in a paper (and the vast majority of our research is of that kind), then there is no possibility to perform this research, since the data used for the previous work is inaccessible. Some people claim that it would be sufficient to describe the major characteristics of the data used, so that other researchers could set up a similar test collection to perform that work. However, we are far away from such an approach—we just don't know the relevant features that determine retrieval quality and thus would allow for repeating IR experiments with different collections.

However, even with the best experimental standards, it is mainly theoretic work that will move our field forward towards an engineering discipline: A good theoretic model has a high explanatory power, and it is based on explicit assumptions. The former allows us to make predictions, and the latter specifies the conditions under which the model is applicable. A more detailed discussion of experimental vs. theoretic work in computer science can be found in [3], where the main conclusion can be summarized in the somewhat paradox statement "There is empirical evidence that the most important contributions to our field are not the experimental ones".

The above discussion has focused on classical, system oriented IR approaches, but it also holds for user-oriented approaches ([4, p. 105] calls for engineering in information seeking) as well as for new research areas in IR [1].

Empirical research is necessary, but it must be accompanied by strong theoretic models. IR evaluations should focus more on answering the *why* questions, and less on *how* to achieve good performance for the test collections at hand. Only this kind of research will put an IR engineer into the position to judge whether or not a specific model is applicable in a given situation, and then use it for making predictions about the properties of the system she is designing.

1. REFERENCES

[1] J. Allan, W. B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2–32, 2012.

[2] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, editors, *CIKM*, pages 601–610. ACM, 2009.

[3] G. Génova. Is computer science truly scientific? *Commun. ACM*, 53(7):37–39, 2010.

[4] P. Ingwersen and K. Järvelin. *The turn: integration of information seeking and retrieval in context.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.