Formalising Evaluation in Retrieval

S. Arafat, C.J van Rijsbergen and J. Jose
University of Glasgow

CoLIS Workshop on Evaluating User Studies in Information Access
Fifth International Conference on Conceptions of Library & Information Science – Context: nature, impact and role. Glasgow 2005

1. Introduction

Traditionally, evaluation methods in the study of information retrieval focus on user studies based on justification from social sciences and psychology. This type of evaluation for retrieval systems has been more empirical in nature than theoretical causing uncertainty and complexity in the definition of research problems and methodology; and in the interpretation of experimental results. It is proposed that the reason for this is the lack of a formal theory unifying the methods of expressing and reasoning about the different search components. Absence of such a theory limits our ability to compare in a sound way the research problems, methods and results in information retrieval. A successful unified theory would be required to amalgamate the different representations of a search component, such as the vector space and probabilistic models at the matching/decision level. It would also require integrating the manners in which a component is described and studied. Ideally this theory would allow search elements to be formally described and reasoned about in relation to one another. The search elements described as the interaction language, interface, decision/matching mechanism and data corpus require to be represented by a common formal theory. We briefly discuss the issues associated with such representation and what it would mean for retrieval evaluation and experimentation were a unified theory be deemed feasible.

2. Evaluation could be more scientific in IR

An important characteristic of *scientific* experiments is that in many cases they can be duplicated exactly in another time and location. This is possible if each of the experimental parameters/conditions can be reproduced. In order to replicate an experiment, the defined parameters need to be well understood. In the natural sciences, experimental parameters often permit detailed, formal specification. If the parameters are well defined then experiments can be conducted with accurate *control* of each parameter/condition. These characteristics of parameters in scientific experiments make computer simulation of experiments feasible creating a potential cost benefit.

One of the main difficulties in simulating an experiment becomes apparent when an experimental factor is not well understood. The experimental setup of a user study in retrieval evaluation therefore contains an inherent complication, the user. If the user was a controllable machine then we could note down user behaviours from one search experiment and duplicate them in another by initialising this machine with certain parameters. The user would then be a *controllable* experimental factor, since an instance of it could be formally defined in terms of behaviours. However, user behaviours are not yet expressed in such a formal manner and instead we are left with brief and informal natural language descriptions, i.e. 'the users are university students with moderate experience in searching'. Ideally one would want to specify exactly, the *character* of each user, so the context of the experimental results can be better understood.

Unfortunately two of the other major elements, the user interface and the user-system interaction language are also usually described informally by natural language. Overall, the way we specify an experiment in IR lacks formal expression. This restricts the capability to formally reason about evaluation results with respect to these search elements. Traditionally, only the document and matching models (Figure 1) have formal specifications in retrieval experiments. In comparison, the method of specification in the sciences, practical physics for example, is formal in more aspects. So we see that the way a retrieval experiment is represented creates ambiguity due to the informal natural language expression of several of its parameters. However, in general IR research, these retrieval elements are represented, written about and thought about in this way, out with experimentation. This is reminiscent of a similar situation from the beginning of the 20th century when Hilbert and von Neumann proposed to axiomatise branches of mathematics and physics [Corry.97], suggesting a systematic approach to expressing theoretical and experimental claims and reasoning about them. The consequence of this drive towards systematic specification was a deeper understanding of the *structure* of the problems being specified and *limitations* of the formal specification system itself; resulting in theorems of computability and incompleteness. The formal specification system that later emerged allowed computers to simulate mathematical models, and anything that could be in turn modelled by the mathematics, such as physical and chemical processes. In order to make the experimentation in IR more like that in the sciences and make good use of simulations, a formal system for expression of search scenarios is necessary.

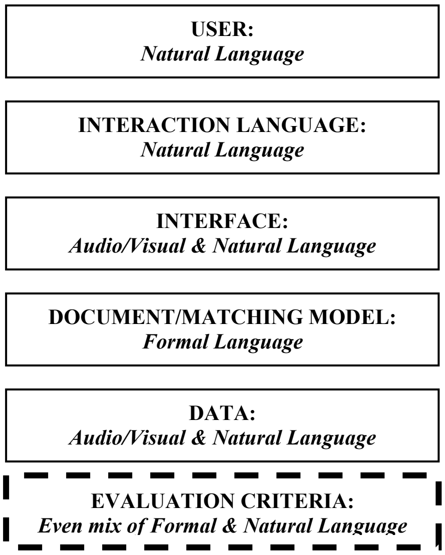

Figure 1: The languages in which retrieval elements are usually expressed

3. Implication of a Unified Specification Language for Retrieval

Formally defining a retrieval process, at first glance, seems implausible due to the complexity of the user element (Figure 1). The mathematical modelling of physics did not begin by attempting to describe the whole universe but by instead describing much smaller phenomena. Hence given a generalised specification language and a method of axiomatisation one could begin by formally representing very specific experiments with very specific user behaviours and gradually building a database of a multitude of possible retrieval scenarios. Although a real user may never be exactly specified, the formalism could reveal facts about the limitations of user simulation just as the theory of computability revealed restrictions inherent in modern day computers. There is also the potential to understand on a deep level when one can do simulation and when they must resort to user studies.

4. Description of Current Research

Our research is based on finding a unifying theory and specification language for retrieval in which all the elements of Figure 1 can be defined. In our theoretical framework we informally define each of these elements in terms of two characteristics - their *representation space* and *method space*. The representation space for an element is defined as the collection of *all* specifications of members of the element. For example, vector spaces are one way to represent the matching and document models; therefore they are part of the representation space for these retrieval elements. The interface can be represented by functional descriptions and diagrams, which form part of its representation space. The method space for each element defines all logical combinations of the specifications in the representation space. These logical combinations are precisely the way in which the vocabulary, that is, the representation space, is used in expressing the element. For example, the part of the method space corresponding to a vector space representation (a part of the representation space) consists of a mixed (natural language with mathematics) description of how to perform a cosine similarity operation and express its interpretation. Alternatively, the representation space corresponds to *syntax* for expressing a retrieval element whereas the method space corresponds to the *semantics* of the representations.

Restating the message of the previous sections in a new way, the problem with IR in terms of this type of characterisation is the multitude of such spaces and there being no way to compare/contrast between them. The aim is first to find for each search element a new representation space, which can accommodate several of the significant representations that are around. A general theory is then required to form a complementary method space to allow reasoning about all such representations in relation to one another. Such a theory would have to agree with the semantics defined by the prior method spaces of individual representations so that the new theoretical constructs exhibit consistent interpretations. The idea is to *map* each of the retrieval elements to a more general space keeping the mapping as *isomorphic* as possible. Following this the task is to find a representation space and method space which can accommodate *all* of the retrieval elements. In the terminology of the previous sections, this final representation space is the unified language and the corresponding method space the unified theory.

On analysis of the matching and document modelling elements we found that a theory of Quantum Mechanics (QM) is able to accommodate the main representation and method spaces, such as probability spaces, vector spaces and logical models which are traditional for specifying the decision and matching elements [van Rijsbergen.04]. It turns out that the theory can also account for many cognitive phenomena and hence accommodate the representation/method spaces describing the semantics of very specific user behaviours [Gabora and Aerts.02]. The user behaviour (a subset of user cognition) is mapped to the QM framework and *inherits* the QM method space. In the new method and representation spaces the user element can be expressed using the vocabulary of the inherited specification language as a quantum-physical process in terms of *quantum state changes, measurements* and *phase-shifts.* Decisions made by a retrieval system, its data representation, and matching models can be expressed using the same set of theoretical constructs and rules when mapped onto the QM framework. The evaluation element once mapped to the QM framework can be interpreted in a manner similar to a *physical measuring device*.

The mapping of elements from their initial method and representation spaces to the generalised representation offered by the QM framework is difficult as in general, the initial spaces are not well defined. Analysis of the large amount of method/representation spaces is required especially for those elements not exhibiting strict formal definitions, such as the natural language elements in Figure 1. The aim in such analysis is to find commonalities in the representation and method spaces of these elements. This would aid in deducing the theoretical constructs for representing the elements in the QM framework, so that the commonalities remain in the new representation. Previous work in [Xie.02] defines some of the common representation and method spaces of the interaction language, interface and some user behaviours. Further research is necessary, aiming to extract a larger collection of these spaces before reasonable mapping can take place.

Mapping retrieval to physics in this way allows inheritance of the formal and experimental strengths discussed in Section 2. Such an ability to formally specify retrieval elements can be used to define retrieval scenarios and their change over the period of a search session. Hypothetical scenarios or simulations can then be formally specified and used to reason about the corresponding simulation results. Simulation in IR has helped with the analysis of specific retrieval scenarios [White et al.04] but there are no guidelines indicating how the current simulation techniques could be adapted to general scenarios. It is proposed that formal specification would provide a deeper understanding of the relation between the simulation parameters and the results. It would exhibit a formal reasoning system, which can be used to decide how simulations are related to one another and the validity of inferring (without running a simulation) the results of one simulation from that of another with the intention of reducing costs. An additional foreseeable advantage of mapping a retrieval simulation onto a quantum physical process is that there are instances where problems (expressed in some QM framework) can be solved more efficiently on a quantum computer than on modern day (classical) computers [Grover.96]. This may benefit retrieval evaluation if quantum computation were to become feasible in the future.

5. Project Proposal

Previously there was an attempt to design software to allow simulation of retrieval scenarios described by our theoretical framework. The SIMINEV (SIMulated INteractive EValuation) project [Arafat et al.04] encountered some major problems:-

1. It became difficult to agree on a feasible architecture for implementation. We could not find a modular, object based design to accommodate the different types of simulation parameters. Proposed software architectures suffered the problem of being too specific where adding new parameters such as new user behaviour, would require extensive modifications to the initial design.
2. There was a lack of understanding of the method and representation spaces of the retrieval elements. We could not find a way to define the simulation parameters which conserved the existing associations between retrieval elements. For example user interaction with a search system is associated with their interactive intentions [Xie.02] and the interface. Interactions, interface and user intention are in general associated in a complex way. Simulation requires at least an initial set of association rules.

The problems are related design of the simulation software requires an adequate understanding of the simulation parameters. Thus, the latter problem requires to first be investigated to gather a database of input parameters for a simulation. The parameters requiring most attention in the definition task are those that are usually informally defined, user behaviours, interfaces and interaction styles. A set of rules defining the association between these parameters will need to be deduced. In light of this a broader project, GENEVA (GENeral EValuation Architecture) is proposed which extends SIMINEV by adding adequately defined input parameters. A simulation experiment using GENEVA would consist of initially defining the retrieval scenario in terms of input parameters in the categories specified in Figure 2. This definition is to be represented in a descriptive natural language form, as objects in the software implementation language and as constructs in our theoretical framework. SIMINEV would then be run with these parameters, corresponding to a simulation of a specific retrieval scenario. The results are specific only to these parameters with generalisation relying on formal deductions within the QM framework or the usual statistical assumptions.

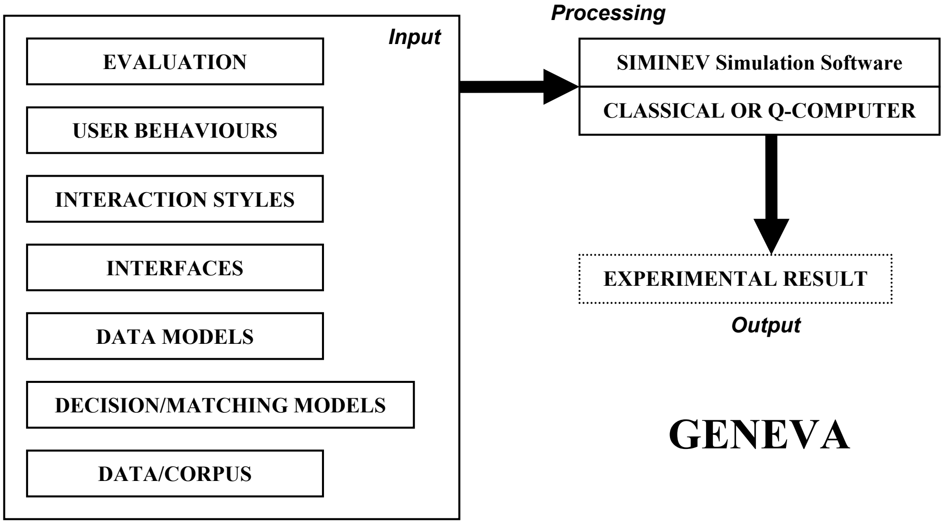

Figure 2. GENEVA (GENeral EValuation Architecture)

GENEVA would initially aim to catalogue comprehensively modern IR research expressing it in terms of the input elements in the above architecture. It would expose *relations* between the search elements so that their representation in our theoretical framework would be faithful to these relations. As above, the process of deriving the input to GENEVA (in Figure 2) requires a systematic categorisation of aspects of IR research and thus is comparable to the drive to axiomatise parts of mathematics and physics. Once there is the input, the simulation software (SIMINEV) can start processing; this is analogous to the creation of computer programs that took place in the mid 20th century to simulate the axiomatised physical theories. A historically significant example is that of the large scale simulation of nuclear detonation models in the Manhattan Project that took place during World War II [Groueff.67].

In conclusion, we proposed that formalising the search evaluation procedure would provide IR research several benefits already present due to the scientific methods, in the natural sciences (see Section 2). The ability to formally describe retrieval scenarios presents opportunities for simulating search experiments/evaluation. Consequently it may be possible to reduce costly user studies in the evaluation process by using simulations which re-use information from previous experiments in a formally justified way.

References

[Corry.97]

L. Corry, "David Hilbert and the Axiomatization of Physics (1894-1905)", *Archive for History of Exact Sciences **51** (1997), 83-198.*

[van Rijsbergen.04]

C. J. van Rijsbergen, “The Geometry of Information Retrieval”. *Cambridge University Press (2004)*

[Gabora and Aerts.02]
L. Gabora and D. Aerts, “Contextualizing concepts using a mathematical generalization of the quantum formalism”. *Journal of Experimental and Theoretical Artificial Intelligence, 14: 327–358, 2002.*

[Grover. 96]
L. K Grover, “A fast quantum mechanical algorithm for database search”, *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing, ACM, New York, pp. 212-19, 1996.*

[Xie.02]
H. Xie**, “**Patterns between interactive intentions and information-seeking strategies”. *Information Processing and Management. 38(1): 55-77 (2002)*

[White et al.04]
R. White, J. Jose, C.J. van Rijsbergen and I. Ruthven, “A Simulated Study of Implicit Feedback Models”. *Proceedings of the 26th Annual European Conference on Information Retrieval (ECIR 2004) Sunderland, United Kingdom, April 2004. 311-326*

[Arafat et al.04]
S. Arafat, C.J van Rijsbergen and J. Jose, “Simulated Interactive Evaluation (SIMINEV)” *Proceedings of the Glasgow-Strathclyde Information Retrieval Workshop, Glasgow, UK (October 2004)*

[Groueff.67]
S. Groueff, “Manhattan Project: The Untold Story of the Making of the Atomic Bomb”. *Published by Little, Brown & Co (Boston, USA), 1967*