A Multi-disciplinary Approach to Interactive Information Retrieval upon Semi-structured Data Sets

Corrado Boscarino
Centrum Wiskunde & Informatica (CWI), Science Park 123,
1098 XG Amsterdam, The Netherlands
corrado@cwi.nl

Abstract

The so called logic and probabilistic views on IR can be reconciled by a unifying framework for IIR. I present a proposal for a PhD research according to a multi-disciplinary perspective and I discuss some of its consequences for IR as a discipline.

Keywords: PhD proposal, Interactive Information Retrieval, dynamic logics

1. INTRODUCTION

Richard Feynman [7] once said that the existence of barriers between disciplines is one of the main obstacles to the progress of science. Conceptual walls are built even inside one single discipline, like between Database-systems (DB) and Information Retrieval (IR) [4] in the burbling pot termed as a whole 'Computer Science', only to find out that they are doing more harm than good to the scientific cause. In the meantime, while the majority of practitioners still believes in the value of a kite-mark for permissible approaches, dwelling in between this received network of disciplines and specialisations is often regarded as an extracurricular activity, which you should not engage with at the expense of the institution that appointed you. In the most fortunate cases it is considered to be a matter of chiefly academic interest, which usually means that only tenured professors may safely pursue that path.

Things start to change when an external event, often the emergence of a technological artefact, confronts us with a paradigm's precinct. For example, the development of nano technologies [14] does not merely improve on the existing engineering practice, but introduces radically different procedures for the manipulation of materials at the atomic scale; the fraction of surface atoms becomes comparable with that of the bulk, the thickness of a layer of material approaches the wavelength of the electronic functions. At this critical scale, material scientists are concerned with quantum effects, chemistry and electronic engineering become deeply intertwined and biology sees the production line as a viable perspective. Practitioners then become aware that landmarks in their own discipline are rough approximations in another discipline, which stay valid until a limit is reached, whether conceptual or technological.

When a larger amount of computing power and advanced data processing applications are accessible by a greater part of the population, as is now the case in Western societies, information becomes a commodity: extraction, processing and distribution of bulk information are not simply extensions in their scope of techniques for accessing structured information repositories, but they are different in kind. Traditional disciplinary boundaries disappear at the scale of indeterminacy reached by modern applications: the amount of disorder in the structure of information repositories and information needs plays in IR the same role as a scaled parameter in physics. Whereas physical laws depend on the scale of parameters such as size, speed or force, methods in IR depend on the amount and kind of structure that informs both the information sources and the queries.

Semi-structured data sets are not less structured deterministic repositories that can be processed by adapting the tools for dealing with structured repositories; their lack of structure blurs the distinctions between IR, DB, and also between these sub-disciplines of computer science, and those fields of philosophy, sociology, anthropology, physics and logic that are concerned with information, its meaning and its use by human agents.

The next section shows that when, information appears to be less and less ordered, and therefore more complex, a successful approach to description, processing and retrieval of information cannot be exhausted within a single discipline. In particular I consider three realms: i) probabilistic models, which realise mappings under the rules of probability theory from the form of the information repository to the form of an information need, ii) semantics, which is considered to either partake only the truth conditions of sentences in a language or more widely to refer to any formal representation of the meaning carried by an informational structure, and iii) user interaction, which may be provisionally termed as any flow of information between a human user and an IR system.

A multitude of disciplines provide valuable insights on all these issues, however applying to IR different conceptual frameworks developed elsewhere would reaffirm the status of an alleged disciplinary essence. In section 3 I shall then introduce my proposal as a demand for disciplines to rethink their theoretical foundations. IR challenges the existing paradigm and we cannot simply resolve to apply different probabilistic models to the problem of evaluating the relevance of a set of documents with respect to a certain query, but we are prompted to address important issues of probability theory through IR. We cannot simply apply theories of meaning to IR, but IR concurs to the development of new theories of meaning. The research proposal that I sketch here has the twofold aim i) to provide a common framework where current approaches to IR, to the description of information flows and to user interaction can be discussed and ii) to provide a design framework where novel applications can be developed, taking advantage of the synergies between different disciplines and improving on the existing theories and practices. While I cannot possibly be exhaustive in the range of applications and on the potentiality of this framework, I will simply present one motivating example of theories, which are known to have been successful in accounting for information exchange related issues, but which are also difficult to reconcile, mainly because they have been developed within the scope of different disciplines.

2. A CURRENT ISSUE IN IIR: SEMANTICS AND PROBABILITY

A large number of insights about the three parts that I regard IIR to be mainly made of are already provided by a variety of disciplines, albeit they are loosely or completely unrelated to each other. In this section I present an issue, which preliminary investigations show that can be tackled with the framework that I aim to. It is related to current debates in the broader field of information access and serves as a starting point for my research.

This exemplary case concerns the long-lasting question whether IR should be concerned with the meaning of information, whatever notion we want to append to this concept, and be about the development of tools designed to ‘understand’ the content of both the documents and the queries and to find suitable mappings between the two. Alternatively, IR can deal only with the outer form of the data, leaving interpretation to the user, provided that enough data are available to make probabilistically responsible statements about relations, further unspecified in their interpretations, between the statistical parameters of the documents and those of the queries. Even at the time of writing, this debate heats the feelings of those involved in information access research [11]; the question whether we should renounce altogether to attempts to model human reasoning only because it is clearly too complicated to be grasped by mathematically elegant expressions is still unsettled.

Without doing too much harm to the multiformity of the different positions, we can cluster a first view on the debate about the relation between semantics and probability around Fuhr’s survey of probabilistic methods for IR [9], essentially based on the application of probabilistic reasoning such as Bayesian methods and on the quantification of relevance as the lumped parameter of

user satisfaction. To the same class of probabilistic methods belongs the language modelling approach [12], which seeks to determine the unique model of the document that generated a query, modulating natural language processing techniques to the task of IR. There have been even attempts to show that the two are rank equivalent [5, pp. 1-10]. Although I do not want to subscribe to this thesis, which has been proved to be in itself problematic [16], the two probabilistic approaches are equivalent in how they put themselves in relation to semantics: since there are a limited number of possibilities to compose the syntactic and lexical units and still produce the same semantic structure, an IR application may simply aim to match those structures, leaving the user to discriminate between the residual subtleties.

Conceptually different are approaches such as that of Nie [15], and before that of Van Rijsbergen [20], who seeks to add semantic awareness to classical probabilistic models. This is at the same time a bridge to IIR in that it enables the employment of different semantic notions for IR purposes, included those that regard meaning to be tightly coupled to human experience. Although this logical approach to IR forms the main anchor to my research, the proposed techniques to convert the output of logic processes to probabilistic statements are still lacking a thorough framework going beyond an *ad hoc* solution for the case of logical implication. The two classes of approaches are seen as competitive only because, as Wong shows [22], IR models are probabilistic implementations of processes of logical inference; when logics are thought of as being mainly concerned with inference we find the logic and the probabilistic view on IR on two different sides if we focus on the implementational layer and to be equivalent at the conceptual level; in both cases it is difficult to design applications where the two views coexist and complement each other.

Relevance may still hold as lumped parameter, but its relation with probabilistic relevance is not easy to determine. The consequence is that relevance can only be characterised in relation to a particular IR system [2], which is conceptually unsatisfactory as we would expect it to be related to the context, the particular query, the previous information gathered by the user, but less to the particular technique that one uses to access information. A rough simplification may lead to affirm that those different approaches to IR are just different estimations of relevance, albeit without a clear understanding of how this parameter is bound to human potentiality. The picture that we get is that of a claim that a certain mathematical expression models human satisfaction without an understanding of what this satisfaction is about.

In order to solve this bottleneck of the design *within* the IR community, without addressing, for example, how anthropologists generalise observations of a local, small scale community, to general cultural theories, very advanced evaluation protocols [21] have been developed. This transfer of theoretical models from design to evaluation does not prevent the closure of the discipline and the elaboration of notions, like that of epistemic uncertainty in [23], specifically targeted to the problem of IR evaluation, although they could be extended to a wider scope. Bringing the human user into the design of IR applications and yet limiting the scope of the theoretical analysis to the boundaries of IR as a discipline also leads to unsatisfactory results. In [11] Fuhr presents a theoretical model for IIR in which he is forced to subscribe a rather strict set of assumptions only in order to let the model be compatible with classic probabilistic IR. In particular concepts such as information need or relevance are left unquestioned.

The few notable exceptions of cross-fertilisation of computer science with theoretical insights from other disciplines remained halfway towards the approach I suggest: they allowed the design of technological artefacts to be inspired by other fields without actively contributing to the other discipline. This is the case of [6], which discuss Merlau-Ponty's phenomenology or Wittgenstein's theory of meaning, both well known and perhaps also slightly outdated, without contributing to the productive field of anthropology of the senses or to that of pragmatics in philosophy of language. IIR should both develop its own theories and communicate them so to trigger advances in the humanities, affording novel ways to do anthropology or philosophy *through* IIR; this is an important benchmark to assess the quality of the solutions developed in such an multi-disciplinary research line.

Before sketching in the next section how a framework that pursues this path further may look like, I shall conclude the present discussion with a methodological remark.[1] It appears that the preconditions to design a multi-disciplinary framework for IIR relies heavily on the expected results of the framework itself; if it does not fall pray of circularity, at least this approach can be facilitated by some bootstrapping procedure. The specialisation of modern research does not allow even a large team to encompass all the different skills that may be relevant to the task. One possible solution to this problem could be to introduce a meta-framework, like that suggested in [1], that provides the key concepts to bridge disciplines. This is, however, still an open issue, which is faced with some regularity in the design of many application in information science. Without this bootstrapping procedure at our disposal, we are then forced to manually select the concepts that, to the best of our knowledge, have been at the core of different disciplines and to use them during the development of the framework.

3. RESEARCH DIRECTION AND CHALLENGES

A probabilistic model of IR can be thought of as a functional block that provides a mathematical formalisation of reasoning under uncertainty and that can be interfaced to other blocks, inducing a hierarchy of different models, each with its specific accessory functionality. A statistical components block attached to the probabilistic model specifies how probabilities should be calculated, one or more logical blocks specify our knowledge about the data sets, the context, the user and so on; logical blocks enter the model through its priors [3], which allow the incorporation of issues that cannot be resolved at the probabilistic level, but which nevertheless may be crucial to the success of the retrieval task. This design scheme overcomes the distinction between the logical and the probabilistic view on IR.

Not yet having an automated method to import other models into the hierarchy, we are then forced to posit a theoretical background, discuss it outside that framework, which still must be designed, and let artefacts based on the framework interact with the probabilistic model; this approach resembles that followed in [13], where, however, little space has been reserved for discussion, such as why the Cognitive framework has been chosen and how it relates to other theories. I also put an additional constraint on the choice of the theoretical approach, that should admit a mathematical representation: either the framework itself must already be expressed in mathematical terms or it must be possible to couch at least its main components in a mathematical language.

A theoretical framework for IIR upon semi-structured data sets should account for the human driven process through which a set of determinately true or determinately false statements can be associated to data sets even when they lack a complete formal structure. Adding the quality of multi-disciplinarity requires also to assess how different theories explain key concepts like meaning, uncertainty or action. Finally, a probabilistic implementation will then assign probabilities to interpretations, given the data set, accessible through a statistical components block, and the procedural and structural information provided through the priors. This section explains, by means of two examples, how different insights on how meaning arises from unstructuredness and interaction, can be integrated towards a common theoretical framework for IIR. For each possible theoretical solution I will provide some clues on how a modular design of novel applications can be implemented in practice.[2]

The first example is an extension of classical logic, which is mainly concerned with assertions and with formalising the task of making an inference, that is making explicit some informations already present within a knowledge base, towards a dynamic logic for IIR. Our task requires a notion of meaning as the product of a dynamic process of interaction between the user, whose capacities far exceed that of drawing inferences, and the external world. User-system interaction and IR is then primary with respect to any crystallised information and knowledge representation structure: meaning does not arise at the system or at the subject in isolation, but at the relation the people

[1] I would like to thank the anonymous reviewer, who pointed out this fundamental issue.
[2] The reader without a background in logic and philosophy may want to skip the technical details.

create with other people, objects and ideas through technology. Multi-agent communication, while it is not commonly regarded in terms of inference

> fall[s] squarely within the scope of modern logic, viewed as a general account of information flow. To emphasise the point, asking a question and giving an answer is just as 'logical' as drawing a conclusion! [19]

The input provided to the probabilistic model through its priors is a characterisation of the dynamics of meaning within a retrieval session. Empirical or theoretical arguments may lead to the identification of different informational events that affect both the context and a user's epistemic state, provided, as I already pointed out, that we are able to produce a mathematical formalisation. Let us suppose that we want to enhance the probabilistic model with a formalisation of the event $!\phi$ of asserting that ϕ in such way that every user, who has access to the system knows that ϕ and she is able to update her epistemic state with that information. Obviously, adjusting the priors with this information leads to a modification of the posterior distributions, for that retrieving the information that ϕ is not relevant any more, the probability of a user asking ϕ will be low, and so on.

A logic block that provides this functionality could be based on a public announcement logic PAL [18], eventually enhanced with common knowledge towards a logic PAL-C. The theory has a mathematical form, hence it is implementable, because the language has a model-theoretic semantics defined onto a model $\mathscr{M} = \langle W, R, V\rangle$, where W is a set of possible worlds, R an accessibility relation from the set of users to the powerset of W^2 and a valuation V fixes the interpretation of variables. As consequence of the announcement ϕ the probabilistic model is updated through the recalculation of the priors that follows each update $\mathscr{M}|\phi$ of the logic model by ϕ. The interpretation of the proposition ϕ can vary in the different blocks, which may be attached to the probabilistic model and, in case of a PAL-block, will be $[\![\phi]\!] = \{v \in W | \mathscr{M}, v \models \phi\}$, that is after the announcement that ϕ, which is supposed to be veridical, only worlds where ϕ is the case are further considered in the calculation of the priors.

Other blocks can be designed, which formalise the acquisition of various structural or procedural informations, thereby determining additional constraints to the possible values of the priors. A functional block that formalise our knowledge about how a IR session proceeds, could be based on a logic of interrogation LoI [10], which admits also a model-theoretic interpretation where an interrogative $?\phi$ partitions the set of worlds into subsets with the same truth value for ϕ and an assertion $!\phi$ selects that partition where the ϕ is the case. Blocks based on temporal logics may formalise the temporal dimension of querying the system and many more can be designed.

The second example, which concludes this section, is an application of anthropological philosophy to the task of characterising a human user, who engages in IR and the way she interprets documents or queries, which do not have a well specified structure. This approach is mainly rooted in Badiou's theory of the subject, but admits, as Fraser shows in [8], also alternative descriptions in terms of intuitionistic logic and Kripke models. Both the meaning and the probabilistic models of complex resources are not expressible in closed form, hence the final epistemic state of the subject and the models are infinite. Also in this case we must seek to determine a formalisation that is expressed in a mathematical language in order to guarantee the feasibility of the calculation of the priors.

The suggested representation is in the form of a generic extension, a finite model extended by a generic set. The concrete documents and queries, being always finite are thought of as approximations by forcing conditions, through which statements about the generic extension are reduced to statements about the finite model. An application of Badiou's idea that the subject is mathematical, connecting an event to elements of the generic set leads to overcome the difficulties in defining the meaning of very complicated resources and the probabilistic models that definitely mark them as relevant with respect to a certain query. Following the exposition in [8], the domain of the subject is described by a set of conditions ©, which is at the same time an element and a subset of the fundamental situation S and $\pi_1, \ldots, \pi_n$ are the sets that define the conditions of ©. The goals of a subjective procedure is a generic truth in the form of a 'correct subset', which is

governed by the rule Rd_1, which states that if $\pi_i \in \delta$ then $\forall \pi_j \subseteq \pi_i \Rightarrow \pi_j \in \delta$ and by the rule Rd_2, which states that, between any two conditions π_i, π_j that belong to the correct subset holds that either $\pi_i \subseteq \pi_j$ or $\pi_j \subseteq \pi_i$; a sufficient condition to ensure that the latter *compatibility* relation holds is that $\forall \pi_{i,j} \in \delta,\ \exists \pi_k \in \delta :\ \pi_{i,j} \in \pi_k$.

The next step towards a mapping of the procedural information about the interpreting subject, consists into defining a spread Δ of correct subsets over © and to add indexes to the set δ in order to indicate the position of the subsets onto a partial order. Defining a spread over © amounts to fix a function that implements the two laws $Rd_{1,2}$ that has $\wp(©)$ as domain and whose range is $\{0, 1\}$. We may consider a potentially generic sequence in Δ, which can arise from a free choice sequence that never becomes expressible by an algorithm. As pointed out by Fraser, we should pay much care to regulate the model in such a way that a certain balance can be reached between the need of keeping the subject free on the one hand and to avoid a restricted law-like sequence on the other hand. This requirement can be satisfied by defining a set ♀such that for every law-like sequence λ, $\lambda(n) =$♀$(n) \Rightarrow \exists m :\ \lambda(m) \neq$♀$(m)$; while it does not allow to definitely indicate at an intermediate step of the algorithm, whether the procedure is generic, it keeps at any point the free choice sequence at a distance from the law. The problem in implementing this block is that any truth procedure is inherently anticipatory and there is no empirical gathering technique that can possibly encompass the entire sequence. At this point, Fraser introduces *forcing* as a means to save the intermediate results of the subjective enquiries. The function of the block is to gradually make sense, by proposing suitable hypothesis and gathering the necessary empirical evidence, of the initially unintelligible terms of what Badiou terms the subject-language, through the forcing relation. Since I present these results in the form of a direction for future research, it is far outside the scope of this introduction to present a complete axiomatisation of the logic of the subject. It suffices here to say that the forcing relation bears a similarity with entailment within an intuitionistic setting; in particular, a set π may force a formula $\neg\neg\phi$ of the language, and yet not force ϕ or it is possible for π to force $\neg\phi$ as long as no other condition of the same generic sequence forces ϕ. What it is also not possible to discuss here is the relation, explained in [17], between intuitionistic, and hence subjective, logics and modal logic; this resemblance has already been observed by Badiou in case of negation that may hold until another construction shows that the positive formula is the case.

4. CONCLUSIONS

After making an appeal for multi-disciplinary research in IIR and for the reconciliation of the probabilistical and the logical view on IR, I introduced a framework where functional blocks, obtained by the mathematical formalisation of different theories, are interfaced with a probabilistic model. I sketched two possible designs: one, which applies dynamic logics, yields the formalisation of public announcements and constraints on the IR system, the second one, formalises the multiplicity of the subject by means of two mathematical structures, the generic set and the forcing constructions. The arguments that I presented in favour of the need for a framework to both discuss different views and to develop novel applications in IIR, seen in its widest acceptation as a technological artefact designed to aid humans in collecting information according to certain criteria, are far from comprehensive of all the perspectives under which we can look at this subject. Nevertheless, I believe to have provided enough support to the claim that the approach that I propose accounts for the complexity of IIR and that only by bundling the strengths of logic, probability theory and philosophically sound theories on human agency can we attain a deep understanding of this subject.

REFERENCES

[1] U. C. Beresi, M. Baillie, and I. Ruthven, “Towards the evaluation of literature based discovery,” in *Proceedings of the workshop on Novel Evaluation Methodologies (at ECIR 2008)*, M. Sanderson, M. Braschler, N. Ferro, and J. Gonzalo, Eds., 2008.

[2] P. Borlund, “The concept of relevance in IR,” *Journal of the American Society for Information Science and Technology*, vol. 54, no. 10, pp. 913–925, May 2003. [Online]. Available: http://dx.doi.org/10.1002/asi.10286

[3] C. Boscarino and A. P. de Vries, “Prior information and the determination of event spaces in probabilistic information retrieval models,” in *Proceedings of the 3rd International Conference on Theory of Information Retrieval (ICTIR 09) - Studies in Theory of Information Retrieval*, 2009, to be published.

[4] S. Chaudhuri, R. Ramakrishnan, and G. Weikum, “Integrating DB and IR technologies: What is the sound of one hand clapping?” in *Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR 05)*, M. Stonebraker, G. Weikum, and D. DeWitt, Eds., VLDB Foundation, ACM SIGMOD. Asilomar, CA, USA: VLDB, 2005, pp. 1–12.

[5] B. W. Croft and J. Lafferty, *Language Modeling for Information Retrieval (The Information Retrieval Series)*. Springer, December 1999.

[6] P. Dourish, *Where the Action Is : The Foundations of Embodied Interaction (Bradford Books)*. The MIT Press, September 2004.

[7] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics including Feynman’s Tips on Physics: The Definitive and Extended Edition*. Addison Wesley, August 2005.

[8] Z. Fraser, “The law of the subject: Alain Badiou, Luitzen Brouwer and the Kripkean analyses of forcing and the Heyting calculus,” in *The Praxis of Alain Badiou*, P. Ashton, A. Bartlett, and J. Clemens, Eds. Elsevier, 2006. [Online]. Available: http://www.re-press.org/content/view/21/38/

[9] N. Fuhr, “Probabilistic models in information retrieval,” *The Computer Journal*, vol. 35, no. 3, pp. 243–255, 1992. [Online]. Available: citeseer.ist.psu.edu/fuhr92probabilistic.html

[10] J. Groenendijk, “The logic of interrogation: classical version,” in *Proceedings of the Ninth Conference on Semantics and Linguistics Theory (SALT-9)*. CLC Publications, 1999.

[11] A. Halevy, P. Norvig, and F. Pereira, “The unreasonable effectiveness of data,” *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[12] D. Hiemstra, “Using language models for information retrieval,” Ph.D. dissertation, University of Twente, Enschede, January 2001.

[13] P. Ingwersen and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[14] Madia and J. William, “Building the future an atom at a time: Realizing Feynman’s vision,” *Metallurgical and Materials Transactions B*, vol. 37, no. 5, pp. 683–696, October 2006.

[15] J.-Y. Nie, “Towards a probabilistic modal logic for semantic-based information retrieval,” in *SIGIR ’92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 1992, pp. 140–151. [Online]. Available: http://dx.doi.org/10.1145/133160.133188

[16] S. Robertson, “On event spaces and probabilistic models in information retrieval,” *Inf. Retr.*, vol. 8, no. 2, pp. 319–329, 2005. [Online]. Available: http://dx.doi.org/10.1007/s10791-005-5665-9

[17] J. van Benthem, “The information in intuitionist logic,” to appear in Synthese special issue on Philosophy of Information edited by Luciano Floridi & Sebastian Sequoiah-Grayson (Oxford). [Online]. Available: http://staff.science.uva.nl/johan/

[18] J. van Benthem, J. van Eijck, and B. Kooi, “Logics of communication and change,” *Inf. Comput.*, vol. 204, no. 11, pp. 1620–1662, 2006.

[19] J. v. van Benthem, “Logic and the flow of information,” University of Amsterdam, Tech. Rep., 1991.

[20] C. J. van Rijsbergen, “A new theoretical framework for information retrieval,” *SIGIR Forum*, vol. 21, no. 1-2, pp. 23–29, 1987.

[21] E. Voorhees, “The philosophy of information retrieval evaluation,” in *In Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*. Springer-Verlag, 2001, pp. 355–370.

[22] S. K. M. Wong and Y. Y. Yao, “On modeling information retrieval with probabilistic inference,” *ACM Trans. Inf. Syst.*, vol. 13, no. 1, pp. 38–68, 1995.

[23] M. Yakici, M. Baillie, I. Ruthven, and F. Crestani, “Modelling epistemic uncertainty in IR evaluation,” in *SIGIR ’07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2007, pp. 769–770.