Structured Information Retrieval and Quantum Theory

B. Piwowarski[1] and M. Lalmas[2]

[1] University of Glasgow
benjamin@bpiwowar.net
[2] University of Glasgow
mounia@acm.org

Abstract. Information Retrieval (IR) systems try to identify documents relevant to user queries, which are representations of user information needs. Interaction, context, and document structure are three important and active themes in IR research. We present how we propose to model the task of Structured IR (SIR) based on a QT inspired framework, with a focus on how to exploit user contextual information and user interaction in the search process.

1 Introduction

Information Retrieval (IR) aims at automatically matching a user's query, usually a set of keywords typed by the user, with a set of relevant documents. Structured IR (SIR) breaks away from the traditional retrieval unit of a document as a single large (text) block and aims at returning document fragments (e.g. a chapter, a section, or a paragraph), instead of whole documents in response to a user query. The structure of the document, whether explicitly provided by a mark-up language (e.g. XML) or derived, is exploited to determine these most relevant document fragments. In addition, SIR users may formulate queries with constraints on the content and on the structure of the units to be retrieved. SIR is believed to be of particular benefit for information repositories containing long documents, or documents covering a wide variety of topics (e.g. books, user manuals, legal documents), where the user's effort to locate relevant content within a document can be reduced by directing them to the most relevant parts of the document.

SIR has been extensively experimented within the INEX evaluation forum[3]. Unfortunately, experimental results so far indicate that, contrary to expectation, exploiting the structure in IR has not led to any significant increase in retrieval performance. One reason seems that models developed for SIR have mainly been adaptation of classical IR models. Even within standard IR, incrementally extending the classical IR models (e.g. adding pseudo/implicit relevance feedback and query expansion components) and adjusting parameters have not led to major improvement in retrieval performance [1].

[3] http://www.inex.otago.ac.nz/

One reason might be that the clues of relevance go beyond topical relevance (i.e. a document is relevant to a query if it is about the topic of the query), even when considering other aspects of the document content (e.g. the document style). More importantly, the *context* (defining the user information need) and the *interaction* (between the user and the IR system) are two important facets that have to be integrated directly into IR models *and* experiments, rather than being a controlled factor. With respect to SIR, structural context (of a document fragment within a document) and interaction (how the user uses structure to navigate within the document when searching) might play an even more important role, perhaps, than in standard IR.

It has been argued that the Quantum Theory (QT) formalism provides new tools for modeling the context and the interaction in IR. We also postulate that QT will allow the modeling of these between the user and the system, and between structured documents parts in SIR.

The paper follows a constructive approach to investigate the construction of a model for (S)IR based on QT. In Section 2, we discuss related works, first the different QT/IR approaches and then the use of structure in IR. In Section 3, we discuss the factors that should be taken into account, and discuss how a suitable model for SIR can be developed in Section 4.

2 Previous works

2.1 QT and IR

In this section we focus on previous attempts of building IR models with the QT formalism. We can distinguish three kind of works – not necessarily incompatible, those (1) adapting an IR model to QT, (2) capturing the user-system interaction and (3) trying to define an adequate space for IR.

Adapting IR models to QT is one of the most direct way to incorporate this formalism into IR. It has two potential benefits. First, it shows that the QT formalism is powerful enough to *at least* encompass those models. Secondly, it might provide an insight on how to modify them in order to leverage the QT "added power". In [2], Van Rijsbergen shows that some classical IR models could be easily translated within the QT formalism. For instance, the vector space model can be easily expressed with the document defining the density and the query as an observable (or vice-versa).

Interestingly, Widdows [3] extended the classical IR vector space model with some concepts brought from QT. Two ideas where put forward. First, non-relevance is equated to the orthogonality in a vector space, and this gives a way to represent negation. Second, the disjunction of queries (q_1 or q_2) is modeled by the subspace spanned by the two subspaces associated with the two queries. Although this adaptation leads to nice results in particular with respect to negation, it does not fundamentally depart from classical IR models.

More recently, Guido et al. adapted [4] the logical imaging formalism [5] to the QT formalism. Logical imaging provides a way to compute a conditional

probability (or an implication) through a non-uniform redistribution of the probability mass. In the classical case, $p(\cdot|A)$ implies that the initial mass associated to $\neg A$ is uniformly redistributed to the space of A. Transposed to IR, it allows for the redistribution of a term probability to an associated term belonging to the query (e.g. synonym). The interesting part of this formalism is the use of a kinematic operator that is an alternative to the standard Shrödinger unitary evolution, and that matches nicely the general imaging framework.

It can be interesting to develop from scratch a framework based on QT, hence avoiding to be restricted by previous standard IR approaches. Such approaches have been attempted in particular to model interaction. The latter has an increasing importance in IR research, as shown by the development of new theoretical frameworks for interaction in IR. Within QT, the work of [6, section 3.7.1.] clearly states one possible operational definition of interaction: *"The [IR system] is a type of oracle which detects a user's question with minimal ed required by the user to express the question, and then provides an answer that maximally satisfies the user"* which in the case of an IR system poses the problem as how to select a set of (part of) documents that satisfies *as much as possible* this requirement. In this work, interaction is modeled as a unitary evolution. Operationalising this framework is still an open question even in the case of "simple" flat documents. Nevertheless we support such a view where system and user are separately modeled, which we discuss in Section 3.

A more practical approach to make use of user interaction was proposed by Melucci [7], who makes the assumption that it is possible to define a Hilbert subspace that contains relevant documents, and that this subspace[4] can be built through user interaction. The QT formalism is involved when, given a document d, the model computes the probability that a document d is within the subspace RS by $\mathrm{p}(RS|d)$ where RS is the constructed subspace.

Somehow orthogonal to the problem of building a model or a space is the question of how to define a space. Usual vector space in IR are spaces where each dimension is associated to one term. Building a concept space (as opposed to a *simple* term space) might be a way to exploit the geometry of the space. Many works [3,8,9,10] have suggested that QT might be adapted to represent concepts. The reasons are two-fold. First, there are interesting connections between orthogonality in the space and the different senses of a word. Second, contextual information could be captured by building up structures where word senses are entangled. For example, knowing in which sense the term "bat" is used might help to identify in which sense "cricket" is used (either animal or sport). These works have the potential to define suitable conceptual spaces for IR.

Two QT based approaches try to build projectors lattices, which in turn can be used to define a Hilbert space. The first one is related to disambiguating words [11]. A lattice of projectors associated to contexts is built, where atomic contexts (e.g. "the *animal* is a tiger") completely determine the context leaving no other possible interpretation for a word (here *animal*). The other work [12]

[4] Melucci names it a context space, but we use this word differently so we do not use his terminology here.

attempts to build a document space where co-occurrence is a central notion. The idea is to characterise texts through *erasing* projectors that only keep words within a given window of a word. While it is not clear how to use these specific frameworks in IR, these approaches are interesting since they define a part of the Hilbert space structure (orthogonality relationships) by the possible observation one could make on the systems represented into that space.

Overall, IR built upon QT foundations is still in an early stage and there are no solid foundations upon which one could develop a sound framework. In section 3 onwards, we discuss some points we deem important for (at least our) future work in that field.

2.2 Structure, psycholinguistics and IR

It is interesting to consider psycholinguistic studies since they might provide an insight on how humans actually use structure. A good summary of current research can be found in [13]. They report that the facilitatory effect of headings in a text is reflected both in the fixations made during the first-pass reading as well as in the later look-backs directed to the topic sentences. At the finer grained level, sentences in the middle of a paragraph can be understood from the structural context of the previous sentence(s), which is not true of the first sentence of a paragraph; transitions between paragraphs and sections thus require more work from the reader. The effectiveness of headings lies in the fact that they provide a mental frame into which upcoming text information may be integrated. At a coarser level, two common hypotheses on why structure facilitates comprehension are stated: (1) facilitate processing of the text topic structure during reading, and (2) readers use text structure to guide text recall (going back to some parts of the text). It hence appears that structure has semantics that could be exploited in IR, because it provides a good way to organize information.

Within the IR community, the use of structure in IR has been extensively studied and evaluated within INEX. Summarizing, structure in IR has been used as a mean to (1) provide more focused material to the user (e.g. return a section of a chapter instead of the whole chapter), (2) specify user constraints on content and structure (e.g. return sections about wine within a chapter about Chile) and (3) provide structural context to a given document part (e.g. a section about jaguar within a book about cars is not the same as within a book about animals). Note that the latter is one possible use of the psycho-linguistics findings, and one that has been shown to improve significantly the performance of SIR systems. Apart from these achievements, structure has not been shown to enhance traditional IR search. We believe new models that use structural context and interaction could make a difference, since they would complement the lack of explicit information about what the user really wants.

3 Factors to consider

To consider interaction in SIR using the QT formalism, our QT-based model should be able to respond to the interaction between the user and the (S)IR

system, which during a search session may include the queries typed and submitted to the system, the clicks users make on links returned by the system, and if available more fine-grained information such as the seen elements (as obtained through the use of an eye tracking tool, for instance). We stress that *all* interactions, including the interaction with the list of results, have to be taken into account in order to build a fully interactive SIR model.

Our QT-based model should also be able to integrate information related to the context of the information need (e.g. previous searches, time, location). The fact that different document fragments may be deemed relevant for a same set of interactions, indicates that relevance is dependent on the search context. This should be captured by the model. Such situations arise for example when the typed query is ambiguous (e.g. "jaguar" as an animal or a car) or when the expertise level of the users are different.

As pointed out in [2], at least two QT features are particularily important to IR. First, the intertwinement of geometry and probabilities, where two distance-wise close vectors representing system states generate almost the space probability distribution on the Hilbert space, and hence the same probabilities of making a given measurement. An example of the usefulness of this principle, is that close-by documents in a term space would imply close-by probabilities of, say, relevance. The second important feature is that measures made on the system might interact with each other in a non standard way, which might prove useful for interactive IR, where for instance a series of observations on the user might change the user state (if we assume that the user state lies in a Hilbert space).

4 A framework for SIR based on QT formalism

We discuss here which space we could be working with, and how it could be constructed for modeling SIR. We first discuss the choice of the representation, and propose to use an information need space. We then discuss how this representation can be used to model interactive IR, and to which extent document structure can be included in the model.

Among the different spaces we could be working with, various choices are possible, but among the most straightforward choice is the topical space [2] – or its approximation, the term space. In such a space, a document is represented by the terms or concepts it contains. Whether this corresponds in QT to a superposition (i.e. a document is a unique combination of terms) or to a mixture of pure term-states is subject to debate, but in both cases a query (or rather the relevance to a query) is an observable, and one can ask the question: "is this document [system] relevant to this query [observable]?". Another kind of questions that can be asked are "is this document [system] about topic X [observable]?".

While this seems to be an intuitive choice, we argue that from a theoretical point of view it is not a sensible choice if we want to use the QT formalism, since it does not exhibit proper quantum properties and does not seem to be adapted to interactive IR.

To uphold the former statement about quantum properties, let us imagine that we have two observable T_A and T_B associated with the observation "this document is about topic A (resp. B)". It can be argued that the two observables interact since the fact that a document is about one topic might influence the fact that it is about another topic. We could even say if we measure T_A, then T_B and eventually T_A, the first measurement of T_A can be different from the second one because asking if the document is about topic B changed the topicality *as perceived by the user*. However, continuing this series of measurement, that is performing $T_A T_B T_A T_B T_A T_B T_A \ldots$, one would expect that the observed values remained the same for both observables T_A and T_B since no new information is brought. This series of measurements cannot happen within QT if no interaction happens, which in this case stems from the fact that users are expected to learn.

In our opinion, these remarks underline two things. First, document topicality is constructive in the sense that any information adds up to previous ones, and this does not match QT measurement in general, since, while measuring, a part of the information is "destructed". Second, we cannot hope to model directly the user perception of topicality as an observable within a document topicality space, since we believe it is a learning process that saturates (i.e. the opinion of the user does not change with further interaction).

4.1 An information need space

Instead we propose the use of an *information need space* where a state, and more generally a density, corresponds to a user information need. Mixed states could naturally be used to model ambiguous information needs, and context/interaction would provide a way to specify what is the actual information need. The density would be pure when the information need is completely determined, as for example when the model can fully predict what are the relevant documents. For an exploratory search (e.g. "I want to learn about Glasgow"), the need density is mixed, whereas for a navigational search (e.g. "I want the University of Glasgow home page") the need density is pure. The relevance of a document (in IR) or document fragment (in SIR) would then be modeled as an observable. This is different from [7], where relevance is modeled as a yes/no observable within a space where documents are the observed systems, and the corresponding subspace is expanded through user interaction.

We think the information need space can model interactive IR since users change their point of view during a search, and relevance, contrarily to topicality, is expected to evolve within a search session [14]. The mechanisms of this change are yet to be understood, but QT could possibly shed a new light on that matter, since this process is not constructive as the document topicality is – users might change their opinion on what they find relevant.

In more details, the information need space could be a tensor product of smaller spaces, each one related to the different dimensions related to the relevance of an information need. A non-exhaustive list of such dimensions would be the topicality, the style (e.g. review, literature, FAQ, etc.), the position in the

structure (e.g. is it a whole book, a section?) and the novelty of the document. Please refer to [15] for a more complete analysis of relevance dimensions.

Without considering context, at the beginning of the search process, the information need space could be seen as a mixed density that corresponds to *all* possible needs, weighed by their probability. What is nice about this is that we could (in theory) provide a list of documents without any interaction and without *any information or interaction* from the user, since it is possible to measure to which extent a document fragment is relevant to an information need density. The context of the search and each interaction would then be extra steps towards the retrieval of relevant information.

Within the various dimensions of relevance, topical relevance is an aspect of the information need that seem to be well adapted to a QT-based model. Let us use an example to illustrate this fact. Consider a user who wants to plan his holidays in Barcelona, and who will be searching for various informations ranging from activities to hotels. Whereas one part of the information need remains untouched (it is about Barcelona in Spain – and not in Venezuela or the Philippines), the other part can drift (from leisure activities to hotels). Interaction through measurement, as described in the next section, would be used to both restrict the subspace to documents about Barcelona in Spain, and to follow the user topical drift from activities to hotels.

4.2 Evolution: Interaction in Information Retrieval?

The evolution of a system is an important topic both in QT and interactive IR. In this section, we study the various forms of evolutions in QT and relate them to our (S)IR.

The first form is measurement. It would account for a partial collapse in the corresponding information need subspaces. An example scenario of interaction would be a user searching for a place to order pizza. At the beginning of the search, the density associated with the information need is not determined and could be a mixture of all possible information needs. The user then types "pizza", which restricts the information need to a given subset of densities and hence to a given subspace of the whole information need space. Knowing that it is 8pm, and that this person is living in a given city would further restrict the density to a smaller subset of densities. More precisely, each new observation (e.g. typed keywords, clicks, time, etc.) would correspond to a possible measurement/projector, and hence to an observable. This integrates nicely within the IR model adapted to QT in [2], since the simplest T would be a projector along the vector representing the keywords defined as in standard IR. In general, the more ambiguous the keywords, the bigger the subspace associated to the projector T.

Note that typed keyword observations can influence more relevance dimensions than the topical one. For instance, if a query contains "review of..." then this is more related to the style of the relevant documents than to their topicality. Linking interactions and measurements would be an iterative process where past interactions could be analysed e.g. in order to compute the exact form of the observables associated to some keywords.

Another possible use of measurement would be to deal with novelty and the related problem of result diversity (that is, how to select a set non redundant pieces of information with respect to a given information need). Documents would be associated to observables within a "knowledge" space for which a user is the system under observation. When a document is read, then the user state would be projected in a subspace that corresponds to a subspace of knowledge where the read document information is known. This process, coupled with the information need specification and drift discussed in 4.1, would be used to build up a list of documents to return to the user.

The second form of evolution would be a unitary one, which describes the evolution of the information need in the absence of interaction. This would be particularly suited to the time observation, since time evolves in a non-interactive manner. Similarly to physics, unitary evolution could also account for the natural evolution of the user's need in the absence of interaction. One possible use would be for example to build up evolution operators using previous user interactions. Again, we can use the holidays in Barcelona example: Users starting to be interested by hotels would turn up to be interested by activities (and vice-versa), leaving the geography-related dimensions untouched.

The third form is through interaction with the environment which in our context is both the user interface and the user memory. This form of evolution should be used when a measurement conducted twice gives two different results. This is the case when, for example, the user interacts with the IR system, and subsequently deems a document to be relevant and latter non relevant, since the user has already read this document. Note that with respect to relevance, if we assume that there is no interaction with the user (i.e. we could use an oracle to tell us that the document is relevant for the current information need state), then we would use standard QT measurement.

To handle the interaction between the user and the SIR system, we would build a user behaviour model. We would define a system space, different from the information need space, where we can represent the current state of the IR system. The state would include information such as which document fragments (or rather hyperlinks to these fragments) are displayed. We would then make the entangled user and system states evolve, taking into account the fact that the user inspects the result list and, in the case of SIR, the behaviour within the document structure, so that to predict which parts of the document collection would be explored by the user. Some part of the interaction would correspond to observations like e.g. when a user clicks on an hyperlink. The result of the interaction ρ can then be measured in this new space, and interaction specific observations like clicks can then be taken into account. The new information need density can be extracted using the partial trace operator, which is useful if we want to reuse this density for new observations and/or predictions.

4.3 Structure

In this section, we discuss how the framework could integrate with structured information.

As discussed in Section 2.2 and in the INEX workshops, structure can help to obtain a better representation of a fragment of text within the document structure whether it be a topical representation, a style or other relevance related dimensions. In our case, to build the topical relevance observable, we could use structure to define the number of dimensions of the associated subspace – ideally, one per topic. An oversimplified example would be to associate each paragraph with a low dimensional subspace of the information need space, and then to build the subspace associated with the section that contains those paragraphs by joining all these subspaces.

Let us note that the bigger (in size) the structural part, the bigger the associated subspace in the information need space, which in turn means that there is a higher chance that a bigger document fragment covers an information need. Consider two document fragments F_1 and F_2, F_2 being included in F_1 (e.g. a paragraph in a section). The projector associated with the relevance of F_1 would "include" (in the sense of inclusion of the projector associated subspace) the subspace associated with the relevance of F_2 (i.e. $F_2 \leq F_1$). Then, if we know that F_2 is relevant to a given query, this would imply that F_1 is also relevant to that query, since the density would be projected into the subspace defined by the projector for F_2, and this subspace is included into the one of the projector for F_1. Deciding which of F_1 or F_2 is better for the user is a matter of user behaviour modeling, as discussed at the end of the previous section.

This nesting property of document fragments also implies that it is not only necessary to find a fragment that covers (*exhaustivity*) the information need, but this fragment has also to be specific to the information need. In order to achieve this, we could build an observable who would measure the percentage of the fragment that deals with the topic of interest. It is relatively easy to build such an observable, since the subspace it spans corresponds to the subspace spanned by the relevance observable associated to the document fragment, but in this case the *specificity* is not a projection observable as the *exhaustivity* is. Both exhaustivity and specificity dimensions are being used in INEX relevance assessments done by human judges, and could be used to compare the output of the algorithms producing the two observables with the values set by the judges.

5 Conclusion

In this paper, we have sketched how contextual and interactive SIR can be modeled borrowing ideas from QT, by defining a space where the user information needs would evolve according to their interactions with the retrieval system. We proposed to use an information need space, as opposed to the standard topical space, as it seems to be better adapted to both IR (allowing interaction) and QT (leveraging a part of the QT framework potential). We briefly described how our information need space, emphasising the fact that it should capture various relevance dimensions beside topical relevance. We then discussed how interaction could be modeled with this representation, and how it would be possible to model the document structure dimension (i.e. what document fragment granu-

larity to return – a paragraph, a section, etc.). While there are still many details to be set in order to get an operational system, we believe this path would allow to capture faithfully the complexity of the search process in SIR.

Acknowledgments This research was supported by an Engineering and Physical Sciences Research Council grant (Grant Number EP/F015984/2).

References

1. Jones, K.S.: What's the value of trec: is there a gap to jump or a chasm to bridge? SIGIR Forum **40**(1) (2006) 10–20
2. van Rijsbergen, C.J.: The Geometry of Information Retrieval. Cambridge University Press, New York, NY, USA (2004)
3. Widdows, D.: Geometry and Meaning. Volume 172 of CLSI Lectures notes. CSLI (2004)
4. Zuccon, G., Azzopardi, L., van Rijsbergen, C.J.: A formalization of logical imaging for information retrieval using quantum theory. In: IEEE proceedings of the 5th International Workshop on Text-based Information Retrieval, IEEE (2008)
5. Crestani, F., van Rijsbergen, C.J.: A study of probability kinematics in information retrieval. ACM Trans. Inf. Syst. **16**(3) (1998) 225–255
6. Arafat, S., van Rijsbergen, C.J.: Quantum theory and the nature of search. In: Proceedings of the First Quantum Interaction Symposium (QI-2007). (2007)
7. Melucci, M.: A basis for information retrieval in context. ACM Trans. Inf. Syst. **26**(3) (June 2008) 1–41
8. Widdows, D.: A mathematical model for context and word-meaning. In: Fourth International and Interdisciplinary Conference on Modeling and Using Context, Stanford, California (June 2003) 369–382
9. Bruza, P., Woods, J.: Quantum collapse in semantic space: Intepreting natural language argumentation. [16]
10. Bruza, P.D., Kitto, K, N.D., McEvoy, C.: Entangling words and meaning. [16] 118–124
11. Aerts, D., Gabora, L.: A theory of concepts and their combinations ii: A hilbert space representation. Kybernetes (2005)
12. Huertas-Rosero, A.F., Azzopardi, L., van Rijsbergen, C.J.: Characterising through erasing: A theoretical framework for representing documents inspired by quantum theory. [16]
13. Hyona, J., Lorch, R.F.: Effects of topic headings on text processing: evidence from adult readers' eye fixation patterns. Learning and Instruction **14**(2) (April 2004) 131–152
14. Xu, Y.: The dynamics of interactive information retrieval behavior, part i: An activity theory perspective. Journal of the American Society for Information Science and Technology **58**(7) (2007) 958–970
15. Saracevic, T.: Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance. J. Am. Soc. Inf. Sci. Technol. **58**(13) (2007) 1915–1933
16. Bruza, P.D., Lawless, W., van Rijsbergen, K., Sofge, D.A., Coecke, B., Clark, S., eds.: Proceedings of the Second Quantum Interaction Symposium (QI-2008). (2008)