How Quantum Theory Is Developing the Field of Information Retrieval

D. Song[1], M. Lalmas[2], C. J. van Rijsbergen[2], I. Frommholz[2], B. Piwowarski[2], J. Wang[1], P. Zhang[1], G. Zuccon[2], P. D. Bruza[4], S. Arafat[2], L. Azzopardi[2], E. Di Buccio[5], A. Huertas-Rosero[2], Y. Hou[6], M. Melucci[5], S. Rüger[3]

[1]The Robert Gordon University, UK; [2]University of Glasgow, UK; [3]The Open University, UK;

[4]Queensland University of Technology, Australia; [5]University of Padua, Italy; [6]Tianjin University, China

Abstract

This position paper provides an overview of work conducted and an outlook of future directions within the field of Information Retrieval (IR) that aims to develop novel models, methods and frameworks inspired by Quantum Theory (QT).

Introduction

The goal of IR is to predict which documents can help users in satisfying their information needs, i.e. to measure the relevance of the documents. Consider this search scenario: "I am looking for information about children activities in Cambridge; this is usually listed at the top of documents; I also want documents containing images. It is raining so I need indoor activities". This information need contains contextual components (local search and weather), which may evolve over time, e.g. after the user has seen some documents. It also contains multimodal components, specifying where the relevant information can be found in the document, and requests text and non-text results. This type of complex but realistic information needs cannot yet be satisfied with today's IR technologies. To address these challenges, we believe that a radically new IR paradigm is needed.

Relationships between formal methods in IR and QT has been shown to exist (van Rijsbergen 2004). In addition, a growing body of literature is supporting the notion that quantum-like phenomena exist in human natural language and text, cognition and decision making, all related to key aspects of the IR process. Corresponding to these quantum-like phenomena are non-classical probabilities that the traditional IR models are insufficient to support. All the evidence suggests that QT provides suitable building blocks for a non-classical approach that could address these challenges.

An on-going international collaboration, through the UK Engineering and Physical Sciences Research Council funded Renaissance project[1], is attempting to develop novel IR models, methods and frameworks inspired by QT. Building upon the Renaissance project, a new EU funded project, QONTEXT, through the Marie-Curie International Research Staff Exchange Scheme, has recently started, aiming to consolidate and extend this collaboration.

Copyright © 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]http://renaissance.dcs.gla.ac.uk/start

Research Hypothesis and Objectives

We hypothesise that QT provides innovation and inspiration into circumventing the emerging context and multimodality issues in IR. Specifically, our main objectives are:

- Development of a generic QT-based paradigm for IR, with special focus on three key themes: (1) **Frameworks:** general frameworks and operational methods for contextual and multimodal IR; (2) **Spaces:** geometrical representation and characterisation of context through semantic spaces; (3) **Interferences:** the interferences among documents, topics and user's cognitive status in contextual relevance measurement process.
- Implementation, application and evaluation of the QT-based IR methods to suitable IR tasks, such as ad-hoc retrieval, interactive retrieval, and multimedia retrieval.

In the following sections, we report the progress achieved in these three key themes.

Frameworks Theme

Interactive IR Framework

The probabilistic formalism of QT provides a sound basis for building a principled interactive IR framework, as proposed in (Piwowarski et al. 2010). In this framework, events, e.g. document relevance, correspond to subspaces, in an "information need space". These provide a means to compute the probability of any event, and to update this probability when the event is realised, e.g. a document is judged relevant. These updates, capturing interaction, are performed on density operators that represent the state of a user's information need. After each interaction, a new density is computed and then used to re-rank documents. A main challenge is the construction of appropriate subspaces and initial density operators. A methodology to construct those from a document collection has been proposed. Current developments include a structured query language to compute the initial information need density, experiments with interaction, multimedia, and polyrepresentation, and the validation of the proposed construction methodology.

Polyrepresentation Framework

Polyrepresentation for IR postulates that a document can be characterised by several different representations. In cross-media retrieval, a document can be represented by textual

and visual features. A document can be looked from different viewpoints, e.g. its authors, its content, user-given reviews, ratings. Each of these properties characterises the document with respect to different representations. If there is a ranking w.r.t. each representation, the intersection of it (i.e. documents appearing in all rankings) is called the "cognitive overlap". Polyrepresentation suggests that relevant documents are at the cognitive overlap of the documents representations. Our aim is to support polyrepresentation in a QT-inspired interactive IR framework, as proposed in (Frommholz et al. 2010). Each distinct representation is described within a Hilbert space, similar to the aforementioned information need space. We define a product Hilbert space of the single representation spaces. Instead of combining the ranking scores independently from different representations, states in this space may be correlated or non-separable due to interrelations between representations. The ranking scores are then computed in a way analogous to the expected outcome of a quantum measurement (Wang, Song, and Kaliciak 2010). Initial experiment has shown a promising performance of the approach in image retrieval.

Logical Imaging

Logic-based models played a fundamental role in the development of IR research. van Rijsbergen (2004) provided the basis for describing logic-based IR models within a quantum formalism, and in particular within Hilbert spaces. Zuccon, Azzopardi and van Rijsbergen (2008) proposed a concrete formulation of a well-known logic-based IR technique, i.e. Logical Imaging, in terms of QT. Logical Imaging (LI) is a technique for updating the probability that a document is relevant to a query by exploiting inter-document term relationships. The new formalisation of LI in terms of QT develops from an analogy between states of a quantum system and terms in documents. In particular, the dynamics of a physical system is associated with the kinematics of probabilities generated by LI. By placing LI within the quantum framework, a mathematical basis is provided to capture, model and use the contextual information associated with a term to understand its meaning in a specific context. The latter was not dealt with in the original LI model.

Quantum Based Measurements for Documents

The problem of representing text documents within an IR system is formulated as an analogy to that of representing the quantum states of a physical system. In (Huertas-Rosero, Azzopardi, and van Rijsbergen 2008), lexical measurements of text are proposed as a way of representing documents which are akin to physical measurements on quantum states. The representation of the text is only known after measurements have been made, because the process of measuring may destroy parts of the text. Thus, the document is characterised through erasure. We define the Selective Erasers as a model of lexical measurement and explore ways of using them to obtain information from documents. Amongst these characteristics are those suitable for using as term weights, based on e.g. occurrence frequencies and distances between occurrences. Erasers are explored as ways of obtaining information about how one term is used with respect to another, like various co-occurrence quantification methods, and distances between neighbouring occurrences. Mathematical structures like lattices and algebras can be defined to generate composite operations able to capture semantic information from text (Huertas-Rosero, Azzopardi, and van Rijsbergen 2009). This research has produced a mathematical foundation for a quantum-like representation of text and provides a basis for indexing and retrieval within a QT-inspired IR system.

Search Simulation Framework

In addition to a mathematical framework, QT provides abstract concepts describing the physical world that may be suitable metaphors for describing IR phenomena. Following this premise, we studied the application of QT concepts to address some foundational issues in search (Arafat and van Rijsbergen 2007). Initially, the problem was to define the concept and process of search and to differentiate it from other processes. This then led to forming a model of search that spurred investigation of problems related to incorporating models of user cognition and interaction within this search model. One conclusion from this work is that, in order to fully benefit from the formalisation tools from QT, a methodological shift in the approach to evaluation in IR is required. Specifically, interactive search scenario simulation would require to be considered. The implications of such a shift are currently under investigation.

Spaces Theme

Context Spaces

Capturing context requires exploiting different sources of evidence involved in the IR process. These sources are the properties of the different entities involved when retrieving and accessing information, where examples of entities include document, task, user, or location. To exploit the variety of entities and sources, it is necessary to model the relationships existing between the entities and those existing between the properties of the entities. Such relationships are themselves possible sources that can be used to predict relevance. In (Di Buccio, Lalmas, and Melucci), we proposed a methodology that supports the design of an IR system able to model in a uniform way the properties of the entities involved, the properties of their relationships and the relationships between the different properties. Sources and relationships are modeled and then exploited through a geometric framework (Melucci 2008), which provides a uniform and concrete representation in terms of vector subspaces. For instance, the degree to which a modeled source occurs in a document can be measured as the distance between the vector representation of the document and the subspace modeling the source(s) spanned by the vector space basis. This motivates the trace-based function proposed in (van Rijsbergen 2004). Using trace in IR, and in particular the density operators, provides the means to establish a link between geometry and probability in vector spaces.

Semantic Subspaces

A major IR challenge is how to represent documents to support the retrieval of those relevant to a user query without retrieving documents that are instead irrelevant. In the vector space model for IR, a document is represented as a vector whose components indicate the (weighted) presence or absence of a term in the document itself. Documents are then matched against a user query, represented as a vector as well, by means of the inner product. Zuccon, Azzopardi and van Rijsbergen (2009b) proposed an alternative approach: documents are represented as subspaces, which encode the semantic of the document itself. However, assuming that a user picks a document, and thus its associated subspace, to represent the query, how could documents be compared to discriminate between relevant and irrelevant? This problem is reformulated as how to distinguish preparations of different quantum systems, where preparations are represented by semantic (sub)spaces. With this in mind, the subspace distance is introduced. Empirical experiments showed that the subspace representation of documents together with the subspace distance provide a technique for separating relevant documents from irrelevant ones.

Non-separability in Human Semantic Space A theory and associated experimental apparatus has been developed for testing whether concept combinations are non-separable in human memory in (Bruza et al. 2009). For instance, the term "bat" can be modelled in a two dimensional vector space as a vector that expresses the intuition that "bat" is a linear combination of two possible senses e.g. "sport" and "animal" (which correspond to vector space basis). The same can be said for the term "boxer" which also has two senses, one dealing with the sport of boxing and the other relating to the breed of dog. As both terms can be represented as a superposition between the basis states corresponding to "sport" and "animal", this opens the possibility to represent the concept combination "boxer bat" following the quantum formalism. In QT, interacting systems are formalised via a tensor product of the individual systems, representing the concept combination "boxer bat" as the outer product of the two vectors. The state of "boxer bat" is decomposable, if it can be established as a product of the state "boxer" with the state "bat". Such a decomposable state of the interacting systems is termed as being separable. We are interested in the non-separability of concept combinations. This has been investigated using a basic quantum-like model of concept combinations. An empirical framework for testing whether such non-separable combinations actually manifest in cognition and some initial results for bi-ambiguous concept combinations are given in (Bruza et al. 2010). Thus far, preliminary results suggest many concept combinations are "pseudo-classically" non-separable meaning the combination cannot be modelled by probability distributions across the senses of the individual words. There are some concept combinations, however, which appear quantum-like entangled but more empirical experimentation is needed to verify this. The non-separability of concept combinations suggests that for an IR system to effectively represent concepts, the representations used should be non-compositional. For example, this rules out using mixture models to represent concept combinations as such models are clearly compositional.

Deriving Pure High-order Semantic Associations from Semantic Spaces

The classical bag-of-words vector space model (VSM) fails to capture high-order semantic associations (interactions) between terms. Initial evidence has shown that the phenomenon of quantum-like entanglement (e.g. as non-separable associations) exists in a semantic space and can potentially play a crucial role in determining the embedded semantics. We propose to consider *pure* high-order interactions (e.g. the terms "invasion", "Napoleon" and "Spain" form a high-level semantic entity) that cannot be reduced to the compositional effect of lower-order ones (e.g. the co-occurrence of "invasion" and "Napoleon" and the co-occurrence of "Spain" and "Napoleon"), as an indicator of non-separable high-level semantic entities. To characterize the intrinsic order of interactions and distinguish pure high-order interactions from lower-order ones, we developed a set of methods (Hou and Song 2009), upon which we propose an extended vector space model (EVSM) that involves context-sensitive high-order information and aims at characterizing high-level retrieval contexts. Compared with the direct incorporation of classical statistical dependence in VSM, our methods have proved mathematically more rigorous. An extensive empirical evaluation is on-going.

Interference Theme

Interference in Document Ranking

Document ranking in response to a user query is usually produced according to the Probability Ranking Principle (PRP), i.e. by ordering the documents according to decreasing probability of relevance. Documents relevance judgements are assumed to be independent from other documents. This is an unreal assumption. In (Zuccon, Azzopardi, and van Rijsbergen 2009a), an analogy between the document ranking scenario in IR and the double slit experiment has been suggested. Through this analogy, it was shown that the PRP resorts to model the ranking scenario as Kolmogorovian probability theory models the double slit experiment. As it is known from empirical observations, the Kolmogorovian predictions for the double slit experiment are inadequate, while those obtained through quantum probability theory correctly reassemble the empirical observations of the physical experiment. This suggests to model the document ranking scenario in IR by adopting quantum probability theory. The resultant is a novel quantum probability ranking principle (QPRP), where interference plays a key role, representing interdependent document relevance. QPRP is able to model situations where a document relevance assessment is influenced by other documents. Empirical validations have consolidated QPRP as the state of the art criteria for ranking documents under interdependent document relevance.

Cognitive Interference

The concept of interference is applied to relevance assessments, aiming to explain and predict user behaviour rather

than the utility of a particular document ranking.

Cognitive Interference in Document Relevance Measurement Cognitive interference occurs when users change their relevance measurement result for one document after they have measured other documents. For example, if users only measure document d_0, they may think it is highly relevant. However, after having viewed another document which is more relevant than d_0, they may now regard d_0 as partially relevant only. We adopt the probabilistic automaton (PA) and its quantum counterpart, quantum finite automaton (QFA), to represent the transition of the measurement states (the relevance degrees of the document judged by users) and dynamically model the cognitive interference of users when they are judging a list of documents (Zhang et al. 2010). The research is inspired by recent work from cognitive science, where QT is employed to model interference in human decision probability measurements which leads to the violation of the classic law of total probability. We are investigating this direction through user relevance judgment data collected from a task-based user study. We also aim to extend it to standard IR tasks and experiments.

Experimental Issues on Cognitive Interference and Quantum Probability In (van Rijsbergen 2004) and related literature, it was hypothesised that a QT-based framework and the quantum probability would be suitable for going beyond classical retrieval models. A IR scenario, in which such a framework can be a necessary, was discussed, and corresponding experiments were carried out in (Melucci 2010). The necessity of considering quantum probability stemmed from the experimental observation that the best terms for query expansion have a probability that does not obey classical probability and instead can be defined within a quantum probability function.

Conclusions and Future Work

This paper reports an ongoing international collaborative research agenda in applying Quantum Theory to develop the field of Information Retrieval and the promising progress that has been made. Current research has been focused on three main themes: Frameworks, Spaces and Interference. We have developed novel QT-based IR frameworks that aim to address two challenging issues: context and multimodality. Common to these frameworks are the construction of appropriate subspaces, their interactions and evolutions, and measurement of document relevance. To facilitate the frameworks, the Spaces theme aims at developing a geometrical representation of context, which can be derived from different sources, and characterising context in terms of semantic subspaces, high-order and non-separable concept associations. The Interference theme aims to model and utilise a fundamental quantum-like phenomenon (as a specific type of contextual effect), namely the interference, in the document ranking and relevance judgment process.

Our future research will be focused on integration and evaluation. Due to the high exploratory nature of this line of research, so far we have taken a bottom-up approach. As a result, the proposed methods in the three themes are still largely fragmented. For instance, exploiting meaning in semantic spaces and incorporating interference in document relevance judgement are means to deal with the context issue, whereas combining evidences and polyrepresentation are to deal with the multimodality issue. Integrating them to form a comprehensive model will be the next step to fully realise the potential of QT-based IR. Further, the current experimental results, carried out separately for different methods, albeit promising, have not yet shown the full extent of using QT for IR. To this end, we will develop a common evaluation methodology and platform, where proper cognition and interaction aspects should be taken into account, and carry out more extensive empirical evaluations.

References

Arafat, S., and van Rijsbergen, C. J. 2007. Quantum theory and the nature of search. In *QI '07*, 114–121.

Bruza, P. D.; Kitto, K.; Nelson, D.; and McEvoy, C. 2009. Is there something quantum-like in the human mental lexicon? *J. Math. Psy.* 53:362–377.

Bruza, P. D.; Kitto, K.; Ramm, B.; Sitbon, L.; Blomberg, S.; and Song, D. 2010. Quantum-like non-separability of concept combinations, emergent associates and abduction. *Logic J. of IGPL.*

Di Buccio, E.; Lalmas, M.; and Melucci, M. From entities to geometry: Towards exploiting multiple sources to predict relevance. In *IIR'10*.

Frommholz, I.; Larsen, B.; Piwowarski, B.; Lalmas, M.; Ingwersen, P.; and van Rijsbergen, K. 2010. Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In *IIiX'10)*, 115–124.

Hou, Y., and Song, D. 2009. Characterizing pure high-order entanglements in lexical semantic spaces via information geometry. In *QI'09*, 237–250.

Huertas-Rosero, A. F.; Azzopardi, L. A.; and van Rijsbergen, C. J. 2008. Characterising through erasing. In *QI'08*, 160–163.

Huertas-Rosero, A. F.; Azzopardi, L. A.; and van Rijsbergen, C. J. 2009. Eraser lattices and semantic contents. In *QI'09*, 266–275.

Melucci, M. 2008. A basis for information retrieval in context. *ACM TOIS* 26(3):1–41.

Melucci, M. 2010. An investigation of quantum interference in information retrieval. In *IRFC'10*, 136–151.

Piwowarski, B.; Frommholz, I.; Lalmas, M.; and van Rijsbergen, K. 2010. What quantum theory can bring to IR? In *CIKM'10*.

van Rijsbergen, C. J. 2004. *The Geometry of Information Retrieval*. Cambridge Univ. Press.

Wang, J.; Song, D.; and Kaliciak, L. 2010. Tensor product of correlated text and visual features. In *QI'10*.

Zhang, P.; Song, D.; Hou, Y.; Wang, J.; and Bruza, P. D. 2010. Automata modeling for cognitive interference in users' relevance judgment. In *QI'10*.

Zuccon, G.; Azzopardi, L. A.; and van Rijsbergen, C. J. 2008. A formalization of logical imaging for information retrieval using quantum theory. In *DEXA'08*, 3–8.

Zuccon, G.; Azzopardi, L.; and van Rijsbergen, C. J. 2009a. The quantum probability ranking principle for information retrieval. In *ICTIR '09*, 232–240.

Zuccon, G.; Azzopardi, L.; and van Rijsbergen, C. J. 2009b. Semantic spaces: Measuring the distance between different subspaces. In *QI'09*, 225–236.