Modeling the Evolution of Context in Information Retrieval

Emanuele Di Buccio
Department of Information Engineering
University of Padova
Via Gradenigo 6/B, 35131 Padova, ITALY
dibuccio@dei.unipd.it

Abstract

An Information Retrieval (IR) system ranks documents according to their predicted relevance to a formulated query. The prediction depends on the ranking algorithm adopted and on the assumptions about relevance underlying the algorithm. The main assumption is that there is one user, one information need for each query, one location where the user is, and no temporal dimension. But this assumption is unlikely: relevance is context-dependent. Exploiting the context in a way that does not require an high user effort may be effective in IR as suggested for example by Implicit Relevance Feedback techniques. The high number of factors to be considered by these techniques suggests the adoption of a theoretical framework which naturally incorporates multiple sources of evidence. Moreover, the information provided by the context might be a useful source of evidence in order to personalize the results returned to the user. Indeed, the information need arises and evolves in the present and past context of the user. Since the context changes in time, modeling the way in which the context evolves might contribute to achieve personalization.
Starting from some recent reconsiderations of the geometry underlying IR and their contribution to modeling context, in this paper some issues which will be the starting point for my PhD research activity are discussed.

Keywords: Information Retrieval, Personalization

1. INTRODUCTION

Information Retrieval (IR) can be framed as a problem of evidence and prediction [19]. Indeed, the purpose of an IR system is to predict which documents are relevant to any information need of any user. What makes IR difficult is that, “like many other things in life, relevance is relative” [23]. In particular, there are many aspects which affect the information need of a user, and consequently the prediction of relevance. The information need depends on the user, the specific task the user is performing, the place and the time. In other words, relevance depends on context.

Since IR is context dependent, the ranking algorithm should efficiently and effectively exploit the information available from the evidence provided by context for predicting relevance. As different assumptions imply different retrieval models, the preliminary assumptions according to which the algorithm works are fundamental. Making these assumptions explicit is important for understanding which information is exploited in an IR model.

The main assumption is that context does not change in time. But this assumption is unlikely. Let consider, for instance, Relevance Feedback (RF) techniques. The idea underlying RF is that the first retrieval operation can be considered as a “initial query formulation” [21]. Some initially retrieved items are examined for relevance; then the automatic modification of the query can be performed by the system by using the feedback collected from the user — for instance adding keywords, selecting and marking documents. The modified query can be considered a “refinement” of the initial query. As shown in [9] in the event of the interpretation of RF in Vector Space Model (VSM), such technique can be interpreted as a form of query context change. The

effectiveness of RF suggests an investigation of the role of the evolution of the context in the retrieval process.

RF techniques point out not only that the information provided by the evolution of context should be considered, but also that such information should be involved in a way that does not require an high user effort. Indeed, even if RF has been shown to be effective, users are reluctant to provide explicit relevance information because they do not perceive it as being relevant to the achievement of their information goals. A possible solution is the adoption of techniques which are transparent to the user, that is “implicit”. Implicit Relevance Feedback (IRF) techniques [7] can use different contextual features collected during the interaction between the user and the system in order to suggest query expansion terms, retrieve new search results, or dynamically reorder existing results. One of the source of difficulties of this kind of techniques is the need of combining different sources of evidence, i.e. different contextual features. The complexity of these approaches is one of the reasons for investigating the problem in a principled way, that is for the adoption of a model-based development. One of the benefits of this approach is that all the assumptions are made explicit: this is crucial in modeling context in order to understand which elements of the context are actually considered, and above all in which way the relationship between such elements is modeled.

The research activity of my PhD will be mainly focused on Personalized Search, that is explicitly considering the aspects which affect the relevance judgments of the user in IR. In particular, the way in which the information provided by the context can be used to achieve personalization will be investigated. Since, at least initially, the problem of the personalization of the information access will be faced in a principled way, in Section 2, after that some of the previous works will be briefly reviewed, some recently proposed frameworks will be mentioned. The discussion about these frameworks will continue in Section 3, where some questions arisen during my preliminary investigations of the problem of personalized search and the study of such frameworks are reported. These questions might be a starting point for my research activity, whose main goal is the design and the implementation of a ranking algorithm for personalized search based on techniques, like the IRF’s, which do not require an high user effort.

2. PREVIOUS WORKS

Since IR is context-dependent, the development of an IR system should consider the information provided by the context. In order to develop a context-aware system, the factors involved and the relationship between such factors should be made explicit. Since the context provides information useful to predict relevance, why is such information not included in the design and the development of many IR system? In [20] the author provides possible answers to this question. A first reason for not considering the context is that, at least initially, such choice allowed the development of IR systems to be simplified. Another reason is that most IR systems are developed to satisfy the need of most of the people most of the time. Personalized Search starts from another hypothesis, that is information access should be personalized. The information provided by the context might be a useful source of evidence to achieve personalization.

The reason for investigating this issue is that previously proposed techniques, which involved some contextual information, were shown to be effective. Let consider a well-known technique, that is RF [22]. These techniques are based on the explicit participation of the user: the user assesses if the documents in an initially retrieved set are relevant, or can suggest or select a number of terms in order to refine its query — as already mentioned in Section 1, RF can be interpreted as a form of query context change. The high user effort required by this kind of techniques together with their effectiveness, suggest to find a way to preserve the benefits of RF and remove its burdens. A possible solution are the techniques based on IRF [7], which use information obtained by the interaction with the documents, for instance, to recommend query expansion terms or retrieve new document sets. The information collected by monitoring the interaction with the IR system can be useful to understand the way in which the information need evolves during the information seeking activities. Many of the IRF algorithms use only one contextual feature, for instance display time [26] or clickthrough data [5]. But, as suggested

in [26], ignoring other sources of IRF means that information about other aspects of the search context is lost. This loss may affect the contribution of the considered factors. A framework which allows multiple sources of evidence to be considered seems to be a necessary solution to exploit suitably the context and its contribution. A problem is to find a theoretical framework for describing the complexity of a system which exhibits contextual behavior and allows the contributions of these factors to be suitably combined. The heuristic-based development can be a possible strategy to address the problem. The latter approach is not negative in itself, but the theoretical framework is to be preferred because “all the assumptions are made explicit and can be reconsidered and refined independently of the particular retrieval algorithms” [24]. Let consider, for instance, the Probability Ranking Principle (PRP) [18]. The assumptions underlying this principle are explicit. Starting from a reconsideration of the classical PRP assumptions, in [2] a new theoretical framework for Interactive IR (IIR) is proposed. The basic rationale is modeling the evolution of the information need by considering that the user moves between situations. “A situation reflects the system state of the interactive search a user is performing” [2]. In each situation the user has a list of possible choices and a positive decision moves the user to a new situation.

The reconsideration and the extension of previously proposed solutions in order to include contextual factors, as in the case of the PRP, is a possible approach. A radical different approach is the one proposed in [25], whose subject is a complete reconsideration of the geometry underlying IR by suggesting the use of Hilbert’s vector spaces.That work constitutes a first attempt of creating a novel and unified IR theory which will allow the emerging challenge of context-sensitive and multi-modal search to be addressed. The VSM is reconsidered also in [9], where the idea of using a basis of a vector space to represent context is proposed. This work discusses also how to model the evolution of the context by linear transformations from one basis to another. Since these works and the proposed geometry can be suitable approaches to address the mentioned issues, in Section 3.1 some of the ideas proposed in such works will be briefly reviewed.

3. CONSIDERATIONS AND POSSIBLE RESEARCH ISSUES

In this section some questions and some considerations based on my preliminary studies on the problem of Personalized Search are reported. They will be the starting point to investigate how to model the contribution of contextual features in the evolution of the information need.

3.1. Why exploiting the geometry of IR?

The starting point of my PhD research activity will be the reconsideration of the geometry underlying IR proposed in [25, 9] and the investigation of some tools provided by Quantum Mechanics (QM) to model the evolution of the information need.

A first reason to consider these approaches is that, as van Rijsbergen states in [25], “the geometry of the information space is significant and can be exploited to enhance retrieval”. Let consider, for instance, the framework described in [9]. In that work the author assumes that a basis of a vector space is the construct to model context. The underlying interpretation is that “a vector is generated by a basis just as an information object is generated within a context” [14]. This interpretation makes possible to describe that an information object can be generated within different contexts: indeed, a vector can be generated by different basis. This framework was shown to be effective and general enough to include contextual features belonging to different contextual levels, particularly the interaction context [12] and the linguistic context[1] [14].
The latter work presents another contribution, that is the *probability of context*, which is the probability that an information object has been generated by a context. The author shows how the probability function proposed to compute the probability of context might discriminate between relevant and non-relevant documents. This function is a trace-based function inspired by the probability formulation in QM. The latter is one of the possible benefits of the adoption, proposed

[1]Here, the expression “linguistic context”, is used to indicate the “users’ context of meaning when they use a particular query term” [6].

in [25], of Hilbert vector spaces[2], one of the mathematical foundation of QM, as basis for a language to model IR. In particular, the mentioned trace-based function can be explained by using one of the results reported in [25], that is the adoption of the Gleason's Theorem [3] to connect Hilbert space and probability. Indeed, according to van Rijsbergen, is "the way in which the geometric structure is exploited to associate probability with measurements" that may be useful for IR. Giving the intuition underlying this view of measurement, might be useful to understand the motivation for reasoning in a more abstract way, that is using Hilbert's space instead of finite inner product vector space as in the VSM.

In QM a *state space* is associated to any isolated (physical) system, which in particular is a complex vector space with inner product [15]. A *state vector* — a unit vector in the system's state space — completely describes the system. Quantities to be measured, named *observables*, are self-adjoint linear operators[3]. The result of the measurement of an observable is one of the eigenvalues of the operator — or better of the matrix representing the operator — corresponding to the observable, "with a probability that depends on the geometry of the space" [25]. Van Rijsbergen states that, one of the interesting properties of this view of measurement, is that it is general and applicable also to infinite systems. He gives some hints about the possible reason for not limiting the investigation to finite systems, particularly with regard to the images. Since at the present time it is not clear which can be the best representation for images, we cannot exclude that complex numbers and infinite dimensionality might be useful to specify operations on this kind of information objects — for instance complex numbers are needed when Fourier transforms of signals are done.

3.2. Contextual Factors and Evolution of the Information Need

QM and Hilbert's spaces may be an intriguing formalism to describe the evolution of the information need because this framework is intrinsically based on the concept of state of a system and it is general and it may be applicable to several "non quantum" domains [16, 17]. Indeed, the framework does not specify which is the state space of a system and the state vector that describes the state of the considered system. There are several works which investigate the connection between QM and IR [25, 1, 10, 13].
Let consider the interpretation described in [13], where the user-information interaction is interpreted as a complex system. In particular, the user-information interaction is described by a state vector of the product space of the state space which describes the document, or the visit of the document, and the state space that refers to the user. An interesting issue might be investigating the evolution of the state of such system. Reaching this challenging objective requires to find an answer to different questions. Finding these answers can help design a ranking algorithm which is sensitive to the context of the user or of the document.

The first question is about the formalism. An in depth investigation of the Hilbert Spaces will be required to understand if this tool could be useful for the objective of my research activity and, in particular, to model the evolution of the information need. In [9] the author proposes to model the context change as a matrix transformation in the proposed geometrical framework. Until now, this technique has not been used to predict relevance. In QM the evolution of a closed quantum system is described by a unitary transformation, in particular by a unitary operator which depends only on the starting and the final time.

- How can this operator be defined?

The detection of the adequate contextual factors is a fundamental issue to be addressed because the observed data is mapped in the vector space basis which represents the context. Probably,

[2] An Hilbert space is a vector space with inner product that is *complete*. Let denote with $(\mathbf{x}, \mathbf{y})$ the inner product between the vectors $\mathbf{x}$ and $\mathbf{y}$, let $||\mathbf{x}|| = \sqrt{(\mathbf{x}, \mathbf{x})}$ be the norm induced by the inner product on the vector space. An inner product space is *complete* if every Cauchy sequences with respect to the defined norm is convergent. In this paper the Hilbert spaces considered are complex vector spaces with inner product.

[3] A linear operator A on a vector space V is an operator $A : V \to V$ which assigns to every vector $\mathbf{x}$ a vector $A\mathbf{x}$ and for which $\forall \mathbf{x}, \mathbf{y} \in V$ and for all scalars α and β, $A(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha A\mathbf{x} + \beta A\mathbf{y}$. The *adjoint* of A, denoted by A^*, is defined by $(A^*\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A\mathbf{y})$. An operator is *self-adjoint* when $A^* = A$.

such contextual factors might be useful also in the definition of the operator describing the evolution of the context. As previously mentioned, the information about the interaction between the user and the system might be a useful source of evidence. But, since the information need is not only related to the user, but also to the task the user is performing, maybe only a subset of the available evidence should be used to refine the prediction of relevance.

- Which are the contextual factors that most strongly influence user behavior during information seeking-activities?

This issue is important also in the event of multimedia IR: the contextual factors could vary according to the medium of the query. Another issue is that a great number of factors can affect relevance and some of these factors may be hidden. The term “hidden” refers to the fact that some information is not only about the user or only about the documents, but characterizes the interaction between these two poles of IR.

- In which way this information can be discovered?

In [13] the author proposes a methodology for IRF which exploits the concept of entanglement. The state of a composite system is entangled if it cannot be written as a product of states of its component systems; this notion seems to fit the case of such properties that are not proper of the document or of the user, but characterize the interaction between the two. In [13] the Schmidt Decomposition Theorem [15] is adopted to determine if a state is entangled or not. Singular Value Decomposition (SVD), which the Schmidt Decomposition is based on, allows the “most influential” contextual factors to be detected. Another possible solution is proposed in [11], where a statistical framework based on Principal Component Analysis is utilized in order to discover the “hidden contextual factors”. Although these techniques provide a solution to the previous question, other existing techniques will be investigated in order to understand the differences in terms of contributions and efficiency. The mentioned techniques might help to find an answer also to the problem of the detection of the suitable contextual factors, a key issue when IRF techniques are adopted. Finding a solution to this problem will be important also when the theoretical results will be experimentally evaluated — some issues related to the experimentations are reported in Section 3.3. The previous mentioned decomposition techniques, as shown in [11, 13], might be a starting point to find an answer to the following challenging question:

- How can the different factors involved be suitably combined and used for the prediction of relevance?

Indeed, the purpose is understanding how the contextual information can be exploited to enhance retrieval. Once the adequate contextual factors are identified and the operator to model the evolution of context is defined, since the purpose is not the mere investigation of the existence of a transformation which describes the evolution of the system, but the final aim is the prediction of relevance, the main question will be:

- Does this operator help predict relevance?

In order to answer this question an in depth investigation of the behavior of the transformation should be done both at the theoretical and experimental level. This investigation might be useful to understand how the state of the system changes with time. Another question is if the QM can improve an approach based on the geometric interpretation of the context change proposed in [9].

3.3. Experimental Validation of the Theoretical Results

The evolution of the user-information interaction will not be investigated only at the theoretical level. The obtained theoretical results will be experimentally validated. But different issues raise because of the adoption of the experimental approach.

A first issue, already mentioned in Section 3.2, is the problem of the detection of the “most influential” contextual factors.

Another issue has to be addressed is the problem of the dataset. Indeed, information about the interaction between the user and the documents is required to test if the proposed approach can be useful to predict relevance and to achieve personalization. An example is the dataset utilized in [26], constituted by interaction logs obtained during a longitudinal user study of seven subjects' interaction behaviors over a period of fourteen weeks. The interaction logs store information like display time, number of keystrokes for scrolling a web page, if a page has been saved, bookmarked or printed. Part of my research activity will be focused on the implementation of a tool to collect this kind of information and which can be subsequently integrated with the functionalities of the ranking algorithm obtained by the theoretical investigation.
Another interesting but at the same time troublesome issue, is the choice of an appropriate measure of retrieval effectiveness. Precision and recall generally are based on binary relevance judgments. But, as stated in [4], the overwhelming number of documents the modern large retrieval environments return to the users, suggests the adoption of graded relevance judgments. The latter approach allows for developing IR techniques which identify highly relevant documents: indeed all documents are not of equal relevance to their user. Highly relevant documents should be identified and ranked first for presentation. A possible solution is a generalization of the measures of recall and precision or the development of novel measures based on graded relevance judgments. In [4] several novel measures are proposed, among which the Normalized Discounted Cumulative Gain (NDCG). The latter is devised to be able to handle useful score ranging in a non-binary scale and to make a better use of multi-level judgments than precision. The problem is that the definition of measures like precision and recall is based on prior human judgments. As a consequence, conclusions to such an experiment are very subjective as they are limited to the scope of the test collection and to the context. As pointed out in [1], a formal method for abstracting user behavior will allow to duplicate and verify experiments. The change of perspective due to the adoption of the QM mathematical framework, will affect also the field of evaluation, and provides an instrument to study a formal model which "approximates" the user. The problem of finding an appropriate measure of retrieval effectiveness might be an interesting issue to be addressed in future research.

4. CONCLUSIONS

This paper is focused on some issues concerning personalized search, that is explicitly considering the aspects which affect the relevance judgments of the user in IR. The main objective of this work is describing the motivation for focusing my PhD research activity on this problem, and in particular on a model-based development. There are several issues which are pointed out during my preliminary investigations. The main problem I am going to investigate is the evolution of the state of the system which models the user-information interaction and if the information obtained by the study of the evolution may be useful for the prediction of relevance and to achieve personalization. Some considerations derived from some recent investigations of the geometry of the information space have been reported. In particular, the motivation for starting from a geometry of IR has been explained, also with regard to the possible connection with QM. Although the relationship between QM and IR requires further investigation, some recent results, like the introduction of the probability of context, seem to indicate QM to be promising to model context in IR. Indeed, the quantum approach may be a suitable formalism when the system under study is sufficiently complex and exhibits contextual behavior [8]. Moreover, the tools QM provides in the fields of Statistics and Probability Theory, might be suitable tools to fit the uncertainty which intrinsically characterizes IR.

5. ACKNOWLEDGMENTS

The author is grateful to Massimo Melucci for the suggestions provided for this paper, and to the reviewers for the useful comments.

REFERENCES

[1] Arafat, S. and van Rijsbergen, C.J. (2007) Quantum Theory and the Nature of Search. In *Proceedings of QI 2007*, Stanford, CA, USA, 26–28 March.

[2] Fuhr, N. (2008) A probability ranking principle for interactive information retrieval. *Information Retrieval*, **11**, 251–265.
[3] Hughes, R.I.G. (1989) *The Structure and Interpretation of Quantum Mechanics*. Harward University Press.
[4] Järvelin, K. and Kekäläinen, J. (2002) Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, **20**, 422–446.
[5] Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. (2005) Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR'05*, Salvador, Brazil, 15–19 August, pp. 154–161, ACM Press. New York, NY, USA.
[6] Kelly, D. (2006) Measuring Online Information Seeking Context, Part 1: Background and Method. *Journal of the American Society for Information Science and Technology*, **57**, 1729–1739.
[7] Kelly, D. and Teevan, J. (2003) Implicit feedback for inferring user preference: a bibliography. *ACM SIGIR Forum*, **37**, 18–28.
[8] Kitto, K. (2008) Why quantum theory? In *Proceedings of QI 2008*, Oxford, UK, 26–28 March, pp. 11–18.
[9] Melucci, M. (2005) Context modeling and discovery using vector space bases. In *Proceedings of CIKM'05*, Bremen, Germany, 31 October – 5 November, pp. 808–815.
[10] Melucci, M. (2007) Exploring a mechanics for context aware information retrieval. In *Proceedings of QI 2007*, Stanford, CA, USA, 26–28 March.
[11] Melucci, M. and White, R.W. (2007) Discovering hidden contextual factors for implicit feedback. In *Proceedings of CIR'07*, Roskilde, Denmark, 20–21 August.
[12] Melucci, M. and White, R.W. (2007) Utilizing a geometry of context for enhanced implicit feedback. In *Proceedings of CIKM'07*, Lisbon, Portugal, 6–9 November, pp. 273–282.
[13] Melucci, M. (2008) Towards modeling implicit feedback with quantum entanglement. In *Proceedings of QI 2008*, Oxford, UK, 26–28 March, pp. 154–159.
[14] Melucci, M. (2008) A basis for Information Retrieval in Context. *ACM Transactions on Information Systems*, **26**, 1–41.
[15] Nielsen, M.A. and Chuang, I.L. (2000) *Quantum Computation and Quantum Information*. Cambridge University Press, UK.
[16] Quantum Interaction 2007. http://ir.dcs.gla.ac.uk/qi2007.
[17] Quantum Interaction 2008. http://ir.dcs.gla.ac.uk/qi2008.
[18] Robertson, S.E. (1977) The Probability Ranking Principle in Information Retrieval. *Journal of Documentation*, **33**, 294–304.
[19] Robertson, S.E., Maron, M.E. and Cooper, W.S. (1982) Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development*, **1**, 1–21.
[20] Ruthven, I. (2004) "and this set of words represents the user's context..." In *Proceedings of the SIGIR 2004 Workshop on Information Retrieval in Context*, New York, NY, USA. ACM Press.
[21] Salton, G. and Buckley, C. (1990) Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, **41**, 288–297.
[22] Salton, G. and McGill, M.J. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY.
[23] Saracevich, T. (1975) Relevance: A Review of and a Framework for the Thinking on the Notion in Information Science. *Journal of the American Society for Information Science*, **26**, 321–343.
[24] Teevan, J. and Karger, D.R. (2003) Empirical development of an exponential probabilistic model for text retrieval: using textual analysis to build a better model. In *Proceedings of SIGIR'03*, Toronto, Canada, 28 July – 1 August, pp. 18–25, New York, NY, USA. ACM Press.
[25] van Rijsbergen, C.J. (2004) *The Geometry of Information Retrieval*. Cambridge University Press, UK.
[26] White, R.W. and Kelly, D. (2006) A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance. In *Proceedings of CIKM'06*, Arlington, Virginia, USA, November 5–11, pp. 297–306.