

On the use of Complex Numbers in Quantum Models for Information Retrieval*

G. Zuccon^(†,‡) and B. Piwowarski^(†) and L. Azzopardi
guido@dcs.gla.ac.uk, benjamin@bpiwowar.net, leif@dcs.gla.ac.uk

School of Computing Science
University of Glasgow
Scotland, UK

Abstract. Quantum-inspired models have recently attracted increasing attention in Information Retrieval. An intriguing characteristic of the mathematical framework of quantum theory is the presence of *complex numbers*. However, it is unclear what such numbers could or would actually represent or mean in Information Retrieval. The goal of this paper is to discuss the role of complex numbers within the context of Information Retrieval. First, we introduce how complex numbers are used in quantum probability theory. Then, we examine van Rijsbergen’s proposal of evoking complex valued representations of information objects. We empirically show that such a representation is unlikely to be effective in practice (confuting its usefulness in Information Retrieval). We then explore alternative proposals which may be more successful at realising the power of complex numbers.

1 Introduction

In the recent years, there has been increasing interest around quantum-inspired models for Information Retrieval (IR). An intriguing characteristic of the mathematical framework upon which these models are based is the presence of complex numbers. While traditional models, such as the vector space models, are based on the field of real numbers, quantum models use complex vector spaces (i.e., Hilbert spaces). Complex numbers are one of the key concepts of the mathematical framework of quantum theory. They allow to describe and model phenomena such as interference, outlined in the next section.

How to harness the use of complex numbers in quantum-inspired IR models has been largely ignored, and this is also the case for most quantum-inspired models proposed in disciplines outside Physics, i.e., the so called “Quantum Interaction” research area [2]. There are three main exceptions. In [6], van Rijsbergen only sketched out the use of complex numbers, proposing to store the term frequency and the inverse document frequency respectively in the magnitude r and the phase φ of a complex number $re^{i\varphi}$. However, no further theoretical insight supporting this proposal has been given, and no empirical evaluation has

* Supported by (†) EPSRC Grant number EP/F014384/ and (‡) Zirak s.r.l. (<http://www.zirak.it/>). The authors are thankful to Peter Bruza, Kirsty Kitto, Massimo Melucci and Keith van Rijsbergen for initial discussion on the use of complex numbers, and to the reviewers for their comments.

been performed. In the context of semantic space models, De Vine and Bruza [3] proposed a novel approach for the construction of spaces based on circular holographic representations, where the construction of complex valued vectors plays a fundamental role in preserving the order information in n-grams. However, they do not provide an interpretation of how complex numbers are used. The same observation applies to the quantum probability ranking principle [8] (qPRP), which relies on the notion of interference. Moreover, in qPRP, as the vector space is not explicitly defined, complex numbers are only implicitly used.

In this paper, we first define what complex numbers are useful for in the context of the mathematical framework of quantum theory, i.e., of so-called “quantum probabilities”. We then demonstrate theoretically and empirically that van Rijsbergen’s proposal does not hold, and discuss how complex numbers could be made explicit for the qPRP based model [8] and conclude.

2 Use of Complex Numbers in Quantum Theory

As stated, complex numbers are pervasive throughout the mathematical framework of quantum theory, due to the wave nature of matter. As such, they provide more freedom in terms of (quantum) probability distributions, and it is this degree of freedom that we describe in this section. Given the space constraints, we make bold simplifications for the sake of clarity.

First, we need to define what a *quantum probability* is. In its simplest form, a quantum probability is characterised by a quantum probability distribution and an event, which are respectively defined by the *unit* vectors \mathbf{d} and \mathbf{e} . The probability $q(\mathbf{e}|\mathbf{d})$ of event \mathbf{e} given distribution \mathbf{d} is then $|d \cdot e|^2$, which corresponds to the squared cosine between the two vectors. This relationship shows that vector based IR can be interpreted within quantum probability theory [6].

Let us analyse further the concept of quantum probability, by considering two vectors on a two dimensional space. Specifically, we represent the event as $\mathbf{e} = \sqrt{1/2}(1, 1)^\top$ and the distribution as $\mathbf{d} = \sqrt{1/|1 + e^{i\varphi}|}(1, e^{i\varphi})^\top$, where \mathbf{d} depends on a parameter, i.e. the angle or phase $\varphi \in [0, 2\pi[$, $|\cdot|$ denotes the usual norm of a complex number, and $\sqrt{1/2}$ and $\sqrt{1/|1 + e^{i\varphi}|}$ are the normalising factors that yield unit vectors. Unless $\varphi \in \{0, \pi\}$, \mathbf{d} is expressed by complex numbers with no null imaginary parts. By varying φ between 0 and π , the probability $q(\mathbf{e}|\mathbf{d})$ varies between 1 and 0. Further, an important fact is that multiplying \mathbf{e} and \mathbf{d} by $e^{i\psi}$ would not change the (quantum) probability value, for all $\psi \in \mathbb{R}$. It is the *phase difference* between the components in the vector that is important. In our example, the phase difference between the two components of the vector in \mathbf{d} is φ .

What does this mean in practice? A simple IR example can clarify the situation. If we assume that $\mathbf{e}_a = (1, 0)^\top$ and $\mathbf{e}_b = (0, 1)^\top$ are documents containing word a and b , respectively, then $\mathbf{e} = \sqrt{1/2}(1, 1)^\top$ means that the document contains both words in equal quantities. By varying φ in \mathbf{d} , we can express that a document is relevant if it contains either a or b , but not both (case $\varphi = \pi$), or is relevant if it contains a , b or both. (case $\varphi = 0$). Intermediate values of φ enable smooth transitions from one possibility to the other.

Table 1. Values of MAP for two matching models based respectively on a real-valued and a complex-valued vector space model (\mathbb{R} -VSM and \mathbb{C} -VSM). Statistical significance using a two-tailed paired t-test with $p \ll 0.01$ is indicated by †.

	AP8889	WSJ8792	LA8990	WT2g	WT10g
\mathbb{R} -VSM	.1870	.1789	.1378	.1276	.1038
\mathbb{C} -VSM	.1313†	.0967†	.1146†	.0781†	.0232†

The idea of using the phase difference between words could also be used in the Quantum Information Retrieval framework [5] where, based on quantum probability theory, the term vector space is used to represent both documents and information needs. In this framework, words can *interfere* between each other in the measurement of relevance.

Interestingly, one could interpret the negative numbers (i.e., $\varphi = \pi$) obtained when performing Latent Semantic Analysis [4] through the prism of the quantum formalism: in this case, a basis vector would contain two categories of words that are mutually exclusive, i.e., that generally do not co-occur.

3 Analysis of the Potentials of Complex Numbers for IR

Encoding idf in the Phase. In [6, page 25], van Rijsbergen suggested to use complex numbers as a sort of information storage mechanism, which then has to be transformed at matching time, where instead of associating to each component of the vector space a $\text{tf} \times \text{idf}$ value, it associates $\text{tf} \times e^{i \cdot \text{idf}}$. As this is the only example of complex number usage in van Rijsbergen’s book, let us go beyond its usage as a simple storage scheme, which is not particularly useful in itself, and interpret it directly as a new complex weighting scheme for documents and queries. Note that we normalised the idf so it ranges between 0 and 2π , since these are the extremal values that a phase can take.

From a theoretical point of view, according to section 2, van Rijsbergen’s proposal would mean that if the query contains a word a with a high idf and b with an average idf, then a document would have a high probability of being relevant if it contains either a or b , but not both! This counterintuitive behaviour does not really depend on the mapping between idf and the $[0, 2\pi]$ range.

For completeness, we experimented with the standard vector space model (\mathbb{R} -VSM) and the “complex” VSM (\mathbb{C} -VSM) on a number of TREC collections. Both documents and queries were indexed with the Lemur toolkit (<http://www.lemurproject.org/>), after applying Porter stemming and stop-word removal. Results are reported in Table 1, and show clearly that the encoding of idf in the phase does not perform well, even when compared to the low baseline of the $\text{tf} \times \text{idf}$ weighting scheme.

Complex Numbers in qPRP. The quantum probability ranking principle (qPRP) is a ranking approach alternative to the traditional PRP that implicitly relies on interferences, and hence on complex numbers [8]. The qPRP has been shown to perform better than other alternatives for the diversity task in IR, and hence

it is interesting to make explicit the representation of documents and to uncover the meaning of complex numbers in that case.

Intuitively, a phase difference corresponds to the fact that documents are relevant for the same topic, and their relevance probability should not add up. A possible re-interpretation of the example of Section 2 is as follows. Assume that a (resp. b) corresponds to the fact that document a (resp. b) is relevant. We can see that with a phase difference of $\pi/2$, a ranking containing the documents a and b would have the same probability of being relevant to the user than a ranking containing only a or b .

How to explicitly encode the relevance of documents and to define the probability distribution is still not clear at this stage. However, the previous example shows that it might be possible to build up the document representation by ensuring that documents do exhibit the same interference as the one that was empirically shown to work well (e.g., defined as a function of the cosine between two documents in the standard document vector space [7]).

4 Conclusions

In this paper we argued that since complex numbers play a central role in quantum theory, it is of interest to harness its extended representational power in quantum-inspired IR models. We have outlined the role of complex numbers in quantum probability theory. We have shown that the proposal of [6] does not hold theoretically or empirically. We have however observed that the qPRP, which was shown empirically to perform well, implicitly relies on complex numbers. In this respect, we have identified a promising direction to further explore the application of complex numbers within IR.

References

1. L. Accardi and A. Fedullo. On the Statistical Meaning of Complex Numbers in Quantum Mechanics. *Lettere Al Nuovo Cimento (1971 - 1985)*, 34:161–172, 1982.
2. P. Bruza, D. Sofge, W. F. Lawless, C. J. van Rijsbergen, and M. Klusch, editors. *QI, 3rd Int. Sym., QI 2009. Proc.*, vol. 5494 of *LNCS*. Springer, 2009.
3. L. De Vine and P. Bruza. Semantic Oscillations: Encoding Context and Structure in Complex Valued Holographic Vectors. *QI 2010*, 2010.
4. T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, editors. *Latent Semantic Analysis. Lawrence Erlbaum Associates*, 2007.
5. B. Piwowarski, I. Frommholz, M. Lalmas, and C. J. van Rijsbergen. What can Quantum Theory bring to IR? In *CIKM '10*, pages 59–68, 2010.
6. C. J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, Aug. 2004.
7. G. Zuccon, L. Azzopardi, C. Hauff, and C. J. van Rijsbergen. Estimating Interference in the qPRP for Subtopic Retrieval. In *SIGIR '10*, pages 741–742, 2010.
8. G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. The Quantum Probability Ranking Principle for Information Retrieval. In *ICTIR '09*, 2009.