LEXICAL MEASUREMENTS FOR INFORMATION RETRIEVAL

A QUANTUM APPROACH

by

Álvaro Francisco Huertas-Rosero

Submitted in fulfilment of the requirements for the title of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

Glasgow, September 30, 2010

Abstract

The problem of determining whether a document is about a loosely defined topic is at the core of text Information Retrieval (IR). An automatic IR system should be able to determine if a document is likely to convey information on a topic. In most cases, it has to do it solely based on measurements of the use of terms in the document (lexical measurements). In this work a novel scheme for measuring and representing lexical information from text documents is proposed. This scheme is inspired by the concept of ideal measurement as is described by Quantum Theory (QT). We apply it to Information Retrieval through formal analogies between text processing and physical measurements. The main contribution of this work is the development of a complete mathematical scheme to describe lexical measurements. These measurements encompass current ways of representing text, but also completely new representation schemes for it. For example, this quantum-like representation includes logical features such as non-Boolean behaviour that has been suggested to be a fundamental issue when extracting information from natural language text. This scheme also provides a formal unification of logical, probabilistic and geometric approaches to the IR problem.

From the concepts and structures in this scheme of lexical measurement, and using the principle of uncertain conditional, an "Aboutness Witness" is defined as a transformation that can detect documents that are relevant to a query. Mathematical properties of the Aboutness Witness are described in detail and related to other concepts from Information Retrieval. A practical application of this concept is also developed for ad hoc retrieval tasks, and is evaluated with standard collections. Even though the introduction of the model instantiated here does not lead to substantial performance improvements, it is shown how it can be extended and improved, as well as how it can generate a whole range of radically new models and methodologies. This work opens a number of research possibilities both theoretical and experimental, like new representations for documents in Hilbert spaces or other forms, methodologies for term weighting to be used either within the proposed framework or independently, ways to extend existing methodologies, and a new range of operator-based methods for several tasks in IR.

To

Maria Eugenia

Julio César

Marcela

Maria Antonia

People very close to what I am.

A good poem is a tautology. It expands one word by adding a number which clarify it, thus making a new word which has never before been spoken. The seed-word is always so ordinary that hardly anyone perceives it. Classical odes grow from 'and' or 'because', romantic lyrics from 'but' or 'if'. Immature verses expand a personal pronoun ad nauseam, the greatest works bring glory to a common verb.

Alasdair Gray, "Prometheus" in "Unlikely Stories Mostly", Canongate Books, 1983

Contents

List Of Definitions

Abbreviations and Nomenclature Used

Abreviations

- **IR** Information Retrieval
- **QT** Quantum Theory
- **PRP** Probability Ranking Principle
- **QPRP** Quantum Probability Ranking Principle
- **NNMF** Non-Negative Matrix Factorisation
- **PLSA** Probabilistic Latent Semantic Analysis
- **BOW** Bag Of Words approach
- **LDA** Latent Dirichlet Allocation
- **VSM** Vector Space Model
- **SE** Selective Eraser (plural, **SEs**)
- **POVM** Positive Operator-Valued Measure
- **LSA** Latent Semantic Analysis
- **LSI** Latent Semantic Indexing
- **HAL** Hyperspace Analog of Language
- **AW** Aboutness Witness

Nomenclature

· $\Rightarrow$ Implication. This symbol is reserved for an operative use: it will appear only as a part of definitions and proofs that are part of the work

· $\rightarrow$ Implication. This symbol will be used freely to represent any implication or conditional as an object of study.

· $\square\!\!\rightsquigarrow$ Aboutness relation. For two information units A and B, $A \square\!\!\rightsquigarrow B$ means "A is about B". Its negation is $\not\square\!\!\rightsquigarrow$.

· $E_1 \circ E_2$ Product between two SEs, understood as the result of applying one (E_2) and then the other (E_1)

· $\prod_{i=1}^{\circ N} E_i$ Product of several SEs, from E_1 to E_N. Equivalent to $E_1 \circ E_2 \circ \cdots \circ E_N$

· $\leqslant$ and $\geqslant$ Ordering relations, specially those between SEs.

· $\cap$ Meet (intersection)

· $\cup$ Join (union)

· $\neg$ Complement

· $\in$ Membership relation

· Π Projector, acting on a Hilbert Space

· $\mathcal{H}$ Hilbert Space

· $\mathcal{H}_A \otimes \mathcal{H}_B$ Tensor Product of Hilbert Spaces

· $A \otimes B$ Tensor product. A and B can be vectors, matrices, or in general, tensors

- · $\underset{C}{\rightarrow}$ Implication relation within a collection C
- · $\mathrm{I}(O_1 \underset{C}{\rightarrow} O_2)$ Degree of implication of two operators O_1 and O_2 within a collection C
- · $\underset{C}{\square\!\!\rightsquigarrow}$ Aboutness relation within a collection C
- · $\mathrm{A}(D_1 \underset{C}{\square\!\!\rightsquigarrow} D_2)$ Degree of aboutness between documents D_1 and D_2 within collection C

Acknowledgements

This work was developed under the funding of the European Comission under the contract FP6-027026 K-Space, Fundación Para el Futuro de Colombia COLFUTURO, Renaissance Project "Towards Context-Sensitive Information Retrieval Based on Quantum Theory: With Applications to Cross-Media Search and Structured Document Access" EPSRC EP/F014384/1, Department of Computing Science (now School of Computing Science) of the University of Glasgow and Yahoo! (funds managed by Prof. Mounia Lalmas).

Special thanks to professor Keith van Rijsbergen and Leif Azzopardi for their effective guidance throughout this work. Thanks also to Leonardo Rosero Hurtado, whose financial support at the beginning of the problem was much appreciated. Finally, thanks to all my colleagues in the Group of Information Retrieval at University of Glasgow, for the fruitful and interesting discussion and pleasant environment. In particular, to Guido Zuccon, Benjamin Piwowarski and Sachi Arafat, for the proof-reading, amongst other kinds of help.

Chapter 1

Introduction

1.1 Representation of Documents in IR

Information Retrieval (IR) is a scientific discipline devoted to solving the problem of picking the pieces of information that satisfy an information need. This is accomplished by retrieving information units from large repositories (usually collections of documents) that fulfil that need. Since a large amount of such units is usually involved, an efficient representation of documents stands as a key requirement; information needs must also be represented as queries in a suitable way to be matched with representations of documents [1].

The task of representing documents has always been central to IR. In the first library systems this task was undertaken by human librarians. These specialists determined what a document "is about" based on a set of keywords, and represented it in a classification scheme [2]. This worked reasonably well, until the amounts of information grew too large [3] to be handled in such a manner. Then it became obvious that, in order to work with big amounts of information, it was necessary to automatise the process as much as possible. Formally describing what a document is about was not trivial for humans, and it could be expected that devising an automatic way of doing it is even more challenging; however, automatic schemes proved quite good at the task [4].

The need to assess *aboutness* was an important problem that appears when representing documents, and it made it necessary to develop ways to formalise and extract *meaning* to a certain extent. Some

techniques were developed for automatic indexing that produce good results and are reasonably scalable[1], but there is still the necessity to improve automatic indexing to make it more sensitive to the subtleties of meaning [5], while remaining practically feasible [6].

In this work, meaning is not going to be defined formally; it is a strongly contextual and complex concept, but we will refer to it as it is reflected in the use of language: meaning is what determines the use of terms (within the rules of grammar). Concepts referring to meaning will be called *Semantic*, mostly in the sense of *lexical semantics* [7] having to do basically with features of natural language text that are determined with the subject the text is about.

A common feature of automatic document-representation techniques, is that they all rely on the adoption of an intermediate level of abstraction bridging the raw data and the abstract mathematical representations used by the information systems. It has been stated that a *geometrical* level is particularly appropriate to work as such a bridge [8]. This mediation between raw data and mathematical representation is in no way specific to IR; it can also be found in scientific disciplines including physics. An important problem of Physics is producing a representation of the state of a system from which all the information that can be obtained through a set of measurements can be derived. In particular, Quantum Theory (QT) provides such a representation: the state of a system and the measurements performed upon it are represented by mathematical concepts; the formal characteristics of these concepts reflect those of the experimental settings that are used for their physical study. QT was developed for the study of natural objects (i.e. photons, electrons, etc) that are not observable directly. These objects seemed to resist a representation in terms of Newtonian physics, and some decades after the theory was put together in its first version, von Neumann found that it included a fundamental non-Boolean logical framework [9].

IR, on the other hand, is a science of quite visible artificial objects (i.e. text / documents) [10] whose characteristics are not given by nature, but by their human creators, and even by their human users. Some central concepts in it, like for example *relevance* have a contextual, measurement-dependent nature that vaguely reminds of quantum concepts.

Is there a reason to apply such physical conceptual framework to such artificial objects? in both QT

[1] A procedure is **scalable** if applying it to larger and larger amounts of information, it still requires a reasonably small amount of resources, like memory, time, processing power. Scalability is closely related to algorithmic complexity (which is its purely mathematical aspect) and efficiency.

and IR, the subject of description by the theory can be **measurement** itself, and the information obtained from it, instead of the objects (particles and natural entities, on one hand, or documents and artificial entities, on the other). This can seem like a non-intuitive approach to physics, but it is precisely what is behind Landauer's principle: "information is physical" [11]. With this catch-phrase, Landauer calls attention on the importance of mathematical representations in Physics, and how they are the main characters in any physical theory.

The relationship between concepts in QT and IR has been explored by van Rijsbergen [10], Dominich [12] and others, but there is certainly much to be studied in this area. This study does not aim to provide an extensive exploration of them, but only taking a simple one and turning it into an implementable framework for IR. This analogy will be introduced in the next section and developed all along this thesis.

1.2 An Analogy of concepts in QT and IR

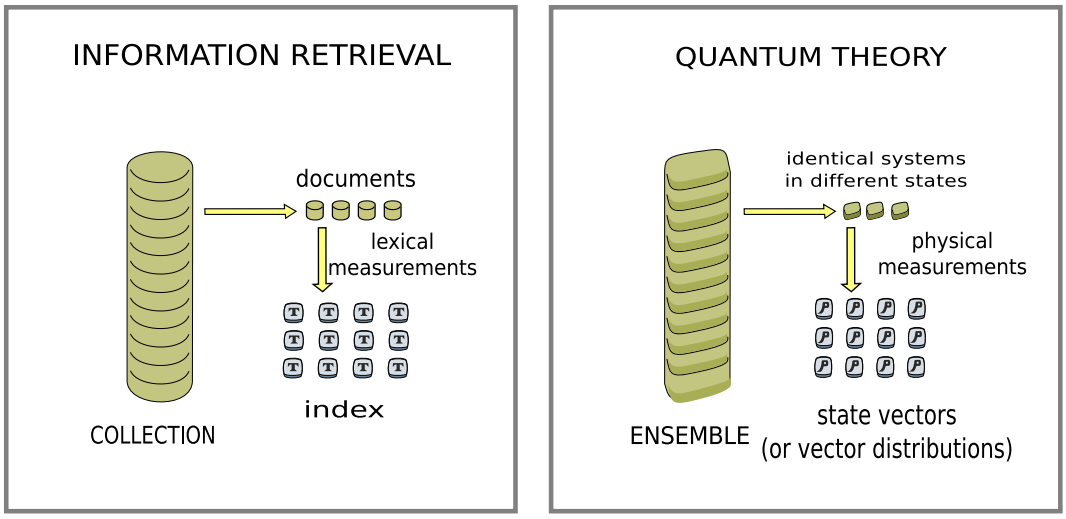

Figure 1.1: Basic analogy between IR and QT

In most text document retrieval tasks, a retrieval system relies on information about terms in order to work: it checks term presence and usage in text to infer similarity or other relations that are needed to assess relevance. This fact suggests the following analogy as a starting point (see figure 1.1):

> Documents (the raw material for IR) can be thought of as states of a physical system (a primary concept in Physics) and their features (such as term occurrences) can be viewed as physical observables to be measured in such system. If a suitable definition of the measurements to be performed on documents is used, then the powerful theoretical machinery of QT can be engaged to represent and use the information obtained. The main contribution of this work is to define suitable lexical measurements which can be performed on text which will form the basis for a document representation scheme.

Historically, quantum-inspired approaches in IR are better understood as coming from mathematical ideas than from physics analogies. Some of these mathematical ideas are:

1. **The algebraic approach**. It has been a recurring theme in the history of sciences that when the knowledge in an area is deep and vast enough, attempts are made to systematise it, and give it a coherent and encompassing structure. Examples of this have been the axiomatisation programs in mathematics and physics, taxonomical schemes in biology and chemistry, or a number of analogous but less successful efforts in social sciences. The most comprehensive systematisation program in IR is possibly Sandor Dominich's *algebraic approach* [12]. According to Dominich:

   > *Abstract structures with an intricate set of relations and no immediate relation to concrete objects can (and do) bear few appeal to part of the IR community, but the very detachedness from anything concrete and worldly is precisely what gives them their flexibility and power*

2. **A logic foundation for IR**. Logics has been at the core of IR methods since the discipline was defined, and at the beginning, *logic* meant *Boolean logic* [13]. However, the necessity of considering *meaning* in an explicit manner made the Boolean concepts problematic, and called for a logic framework beyond Boole, as was pointed out by van Rijsbergen in 1984 [8]. The quest for the appropriate logic concepts (in particular, a *conditional*) able to deal with meaning has led naturally to situation theory [14] and modal logics [15], but the complexity of such approaches have prevented them from being tested properly in IR tasks. More recently, van Rijsbergen also proposes a different approach [10]: Given that propositions about

measurements on a system are intrinsically contextual in QT, the procedures and formalisms used within this theory to represent these measurements and derive predictions from them, can inspire similar methodologies in other areas where context-dependent measurement are involved, like IR.

3. **Vector Space Models (VSM)** Representing documents as elements in a vector space, an early idea in IR, is already something that is close to the representation of states in QT. Salton, Wong and Yang proposed in 1975 to represent document as vectors of features, so they could be compared by distance and similarity vector space measures [16]. Even though the idea involves vector spaces like those used in QT to define vector representation of states, this first VSM were not formulated with any mention or relation to QT. The use of some of the concepts from Quantum Theory to link vector spaces to probabilities was proposed by van Rijsbergen [10]. Widdows proposes in [17] another quantum-like use of vector space representations through a *quantum negation*, and Bruza *et al.* proposes a quantum interpretation of the relation between the meaning of different terms as *spooky activaction at distance* [18], analogous to the way Einstein referred to quantum nonlocal effects (*spooky action at distance*).

4. **Probabilistic Models** The principled approach to IR that has been more successful in IR is probably formulating the problems in terms of probabilities, to allow the use of all the mathematical machinery that has been developed to deal with them. One way of doing it is for example describing occurrences of terms as basic observable events; a probability is computed for them from a sampling on documents. The event of a particular document being relevant, on the other hand, is described as a more complex event whose probability is strongly related to those of term-related (lexical) events. In the decade of 1970, the Binary Independence Retrieval model was proposed [19], based on a simple estimation of the probabilities of occurrences on terms in relevant and non-relevant documents. Recently more complex models, called *discriminative* models have been also proposed, which translate the problem of assessing relevance into a classification, most of them inspired in the Binary Independence Model [20]. Other kind of models that have been developed are *generative* models, which model text as patterns generated by random processes which would differ according to the topic the text is about [21].

This work was strongly inspired by all these mathematical ideas, and can actually be seen as an attempt to apply them in a novel, unified way to IR. Abstract structures that are present in purely algebraic approaches can also be found in this work, naturally arising from the description of measurement. Elements are also given for the interpretation of the involved concepts in terms of logics, probabilities and vector spaces; all these aspects of the description are naturally related in the Quantum description of systems and measurements, and can also be related in the framework proposed in this work.

1.3 Research Questions

The basic analogy relating QT and IR is a very general idea, and the overall objective of this work is not only to fully develop it formally, but also to take it closer to the realm of practical tasks and concrete applications of IR. Measurement will be the basic concept chosen to build upon, and the formalism will be developed from it through addressing five basic research questions:

RQ1. How can basic lexical measurements on documents be defined to match in a very general way the properties of a quantum measurement?

> A *basic lexical measurement* is a procedure to obtain quantitative information about the use of terms in text. This research question will be addressed by defining a procedure performed on text documents that can be mathematically described in a way similar to that for physical measurements.

RQ2. How can basic lexical measurements capture the features of text that convey meaning?

> The use of the newly defined quantum-like measurement will be put to the test first by analising the traditional features of text like term frequencies, but also others that go beyond traditional methods, like burstiness and distances between occurrences. To what point they convey meaning isa question that is not going to be addressed, being a very complex matter of research on its own right; however, simple sanity checks will be shown with a few examples.

RQ3. How to use this approach as a starting point to design better performing IR systems?

This research question will be addressed by defining operations that allow to combine basic lexical measurements in complex operations. These will be again obtained from the analogy with QT, which makes use of a wealth of operations neatly organised in algebraic structures. This question will be addressed in a twofold way: on one hand, simple methods will be developed to obtain and represent information from text documents; on the other hand, an operation on text documents will be built on lexical measurements that can be directly applied to text retrieval.

RQ4. Does the point of view proposed (processing of lexical information as a physical measurement) include existing accounts?

This question will be addressed by formulating the most usual IR approaches in terms of the proposed lexical measurements. This question will also motivate an exploration of relations between existing approaches through the formal elements of the proposed approach.

RQ5. Does the point of view proposed go beyond existing accounts in a fundamental way?

This question will be addressed mainly by focusing on logical aspects of the approach like non-distributivity and the nature of uncertain conditional, that have called for new kinds of model departing from Boolean logic.

1.4 Contributions of this Thesis to the Knowledge in IR

Addressing the research questions that are formulated in last section will produce concrete contributions to the area of IR. They are the following:

- A fully developed scheme for the formal description of lexical measurements that resembles quantum measurements. This can be found in chapter 4.
- A methodology for assessing aboutness based on the principle of uncertain conditional, that makes use of lexical measurements in a way inspired by Quantum Theory.

- An implementable scheme to approach Information Retrieval tasks (in particular, *ad hoc* retrieval) based on simple processing of the proposed lexical measurements.
- Outlines of applications to various IR tasks, as well as hints on how to use them for problems outside this area.
- Preliminary experimental tests of several of the proposed applications, performed and evaluated in standard collections.

1.5 Publications

A number of publications have arisen during the course of this thesis:

1. **"Characterising Through Erasing: A Theoretical Framework for Representing Text Documents, inspired by Quantum Theory "**
   Proceedings 2nd AAAI Quantum Interaction Symposium, College publications, 2007, pages 160–163 (2008)
2. **"Eraser Lattices and Semantic Contents: An Exploration of Semantic Contents in Order Relations between Erasers"**
   Lecture Notes in Artificial Inteligence, vol. 5494, pages 266–275 (2009)
3. **"Selective Erasers: A Theoretical Framework for Representing Documents Inspired by Quantum Theory"**
   2nd BCS IRSG Symposium: Future Directions in Information Access, London, 2008
4. **"Eraser Lattices for Documents and Sets of Documents"**
   Third BCS-IRSG Symposium on Future Directions in Information Access (FDIA 2009) Padua, 2009

1.6 Outline of the thesis

The structure of the thesis is as follows

Chapter 2: Context

Brief description of the problems, tasks, and main methods of the discipline of IR, together with some theoretical concepts that are used in them. This section is divided in five sections:

- Information Retrieval and its Theoretical Basis

  Brief description of the general problems concerning IR, and the ways that have been used to tackle them.

- The Scope of this Work in the Context of IR

  Delimitation of the scope of IR problems with which the current work is concerned, and some reasons to limit it.

- Basic Models in IR

  Aspects of the usual models, methods and approaches of IR that are relevant to the current work are mentioned, with an emphasis on their problematic features and

- Uncertain Conditional as a Unifying Concept

  Definition of Uncertain Conditional with its characteristics and how it is used in IR. His fundamental role in IR, and how it relates to the discussed approaches.

- What Quantum Theory has to offer to IR

  Introduction to some concepts from Quantum Theory that have been proposed for IR, and how they are going to be used in this work.

Chapter 3: Measurement

In this chapter the general problem of measurement is presented in its conceptual and logical aspects, and the advantages of approaching IR from the measurement perspective are explored. This subject is developed in three parts:

- The General Problem of Measurement

The process of measurement is discussed from its mathematical foundations, and the paramount that logic plays in it is described.

- Measurement as Filtering

  A view of measurement based on the filtering of information is presented. Filtering as a selection of the cases where a condition is fulfilled or not, will allow to relate measurement to the logical assessment of propositions such as "condition x is fulfilled", as well as the composition of such propositions with logical connectors (and, or, etc). This will connect to the approach of QT as formulated by von Neumann, highlighting the main differences with the corresponding concepts as would be described by usual Boolean logics. Distributive law is presented as a critical aspect of the logical formulation in which quantum and classical descriptions divert.

- A Logic of Lexical Measurements for IR

  The relevance of the problem of lexical measurement for IR is discussed. Theoretical, and even practical possibilities brought by a change in the conception of Measurement to IR are explored.

Chapter 4: Selective Erasers

The Selective Eraser, basic concept upon which the whole work is built, is presented in this chapter, with a discussion and examination of its properties. Their description is developed in the following sections:

1. Definition
2. How SEs fulfil the laws of measurement outlined in chapter 3
3. Norms from documents: getting numbers out of measurements with SEs
4. Examining text with SEs
5. Order relations between SEs: logics of measurements with SEs.
6. Representation of Documents with lexical measurements

7. Non-Boolean algebra of SEs
8. SEs and probabilities
9. Linear Algebra for SEs
10. Uncertain Implication and SEs
11. Summary

Chapter 5: The Aboutness Witness

In this section we explore how Selective Erasers can we used to assess whether a document is about a topic. Selective Erasers are combined to form a complex transformation called *Aboutnes Witness*, which will be sensitive to semantic contents. The physical analogy that suggests its name is explained, and some of the possibilities for defining it, as well as the properties obtained by each, are explored.

An example: Ad Hoc Retrieval

A practical application of the Aboutness Witness is implemented and tested for a simple *ad hoc* retrieval task, with several standard test collections. We discuss how the proposed approach is equivalent to existing ones, but can generalise them to include features beyond single-term bag-of-word schemes.

Conclusions and Future Perspectives

Finally, in this chapter the whole proposed framework is reviewed, emphasising the new possibilities it brings, and how it goes beyond existing theoretical approaches. The practical applications that have been tried and those that are suggested from preliminary explorations are discussed.

The fundamental character of the work is also examined, together with its place in current IR research and its relations to other state-of-the-art approaches.

Chapter 2

Context Survey

The present work attempts to introduce a novel way of representing lexical measurements which, in turn, leads to a new approach in IR. In this chapter, some of the concepts that are subject to this revision are presented, focusing on the problematic issues that can be recast or even clarified by a new fundamental approach. Some of the theoretical basis of IR are described in section 2.1: definition of the basic problem of IR (subsection 2.1.1), how the user and his or her context can be considered in an implicit way (subsection 2.1.2), and basic components of an IR system (subsection 2.1.3). Since the scope of the basic problem is very wide and makes a formal approach difficult, a simplification of the problem is proposed, by focusing in the aspects that are less user-dependent. One of the aspects of this simplification will be to shift the focus from *relevance* to *aboutness* as the main property to assess in a document. These concepts are both defined in section 2.2, where the properties of aboutness are formulated, to be used later as a sanity check for the methodologies we will devise to assess whether a document is about a topic. In section 2.3, the most important classes of models that have been used in IR are briefly described, with some remarks on how the present approach can unify them formally. The concrete retrieval methodology proposed in this work is also situated in the scheme of IR techniques. In section 2.4 we discuss a concept from logic that has been suggested to underly basic concepts in IR: the *uncertain conditional*. This concept will be addressed also throughout the whole thesis, as one of the formal features of IR that arises naturally from the measurement approach. In section 2.5, some previous works using concepts from QT in IR are briefly described. Finally, in 2.6 the main challenges and tools left by the current state of fundamental research in IR are reviewed.

2.1 Information Retrieval and its Theoretical Basis

Even though in this work we will use analogies between IR and QT concepts, we should be careful to keep their differences clear. IR is not too similar a science to physics. At a methodological level, there are no obvious analogies between these two sciences. IR was developed as a collection of ways of solving concrete problems [22], with occasional theoretical deep explorations on the resulting solutions that give them a sound fundamental basis. A *problem*, for example the indexing of documents in a library, leads to a *solution*: a system of automatic indexing. The solution can then be embedded in a more general *theory*, for example probabilistic retrieval.

A general theory suggesting the nature of IR concepts was usually not the starting point, as it has been the case for physics and other "hard" sciences. This experiment-first way of research has been extremely fruitful: good solutions have been given to most of the initial problems, and the solutions are so good that in early key papers on the field, like [23], methods and concepts can be readily recognised that are still widely used. The way they are implemented and tested also looks quite similar to current research in some cases. However, this work is motivated by the belief that at this point IR would be greatly benefited by a principled approach, going from theory to reality, after a long history of more heuristic approaches going from reality to theory.

The problem-oriented approach to IR has perhaps been a consequence of its applied nature [24], which has modelled it as a fragmented discipline:

> The real-world problem-solving focus of core specialism in IS[1] (e.g. Information Retrieval, Information Behaviour), which disputes about which research problems are most important. The research agenda is often driven by non-academic interests (professional associations, practitioners, governments who want 'useful' research)

Cases of sciences other than IR are known where it has happened otherwise; theory can precede the formulation of the problem itself. This is the case, for example, of superfluid He^4 [25]: a *theory* (quantum boson statistics) leads to the formulation of a *problem* (how quantum fluid and normal fluid coexists?) for which a *solution* is found within the theory (Landau's "second sound" model)].

[1] **IS** stands for Information Science

The problem of coexistence would not even exist, were the concept of the involved phases not introduced by the theory.

As has been stated in chapter 1, analogies to other sciences could help filling the gap of general theories in IR; in particular, analogies to sciences whose theories were built in a very abstract way, aiming to a very general range of application.

2.1.1 The Basic Problem of IR

From its very beginning, IR has been defined as a discipline by a single problem [26], easy to state but very difficult to solve:

> Within a large, possibly unstructured, collection of information objects (documents), retrieve those fulfilling a certain information need of a user.

Defining formally what an *information need* is can be very difficult and is generally not even attempted, so it could be said that IR has been developed as a precise means to an imprecise goal. The complex and problematic nature of this definition arises mostly from the involvement of a human user embedded in a social context trying to accomplish a particular task. In this work, for simplicity, we will avoid dealing directly with user-related issues, by considering only indirectly his or her information needs only as they are formulated in machine-usable **queries**.

Definition 2.1 (Query)
A **query** is a formulation the user has produced about his or her information need as a sequence of terms, that can be directly processed by the system without further human intervention.

We will keep in mind that this represents a contextual and potentially vague information need, but will only address this fact by imposing certain requirements to the logic framework we are using to process the query; however, we will consider the basic IR problem as approachable from two different points of view: that of the user, and that of the system.

2.1.2 System and User: the Two Sides of IR

In the first library systems the user was expected to provide precise and effective descriptions of his or her information need, in terms of a precise code, an index table, keywords selected from a standard set, or some kind of sophisticated but not necessarily simple classification scheme. However, this expectation has proven quite difficult to fulfill; now it is in fact accepted that it should not be the user who adapts him or herself to the system, but the other way around. IR community is making important efforts to acquire a better understanding of the user and his or her behaviour to teach the system how to adapt to it [27, chapter 3].

At the same time, the explosive growth of the IR knowledge in the last two decades has also produced a wealth of different specifically defined IR tasks that allow researches to limit the influence of the specific circumstances of the user in retrieval scenarios, and study in depth the entities and phenomena that are only involved in the processing of data that has fed to the retrieval system.

These two coexisting, complementary trends reflect two sides of IR that reinforce one another, which we will call the system-oriented and user-oriented approach.

Definition 2.2 (System-Oriented approach)

The **System-Oriented approach** considers collections, documents, queries and their representations, as well as processes involving them, as the object of study, and takes the user simply as a given entity that manifests itself by producing queries and requiring certain characteristics (aspects of relevance) from the retrieved documents. The user's nature, behaviour and context is considered, just not as an object of study, but as part of the definition of a given problem.

Definition 2.3 (User-Oriented approach)

The **User-Oriented approach** considers the user as the object of study, as well as his or her behaviour as an agent trying to perform a task, the way he or she interacts with the system, the context relevant to the task and user, etc. It can consider the collection, documents, hardware, methods, etc. as part of the problem under study, but, for simplicity, takes most of its elements and characteristics as given.

In this work, we will adopt a system-oriented approach. This means that we are considering the influence of the user indirectly, as acting through the following items:

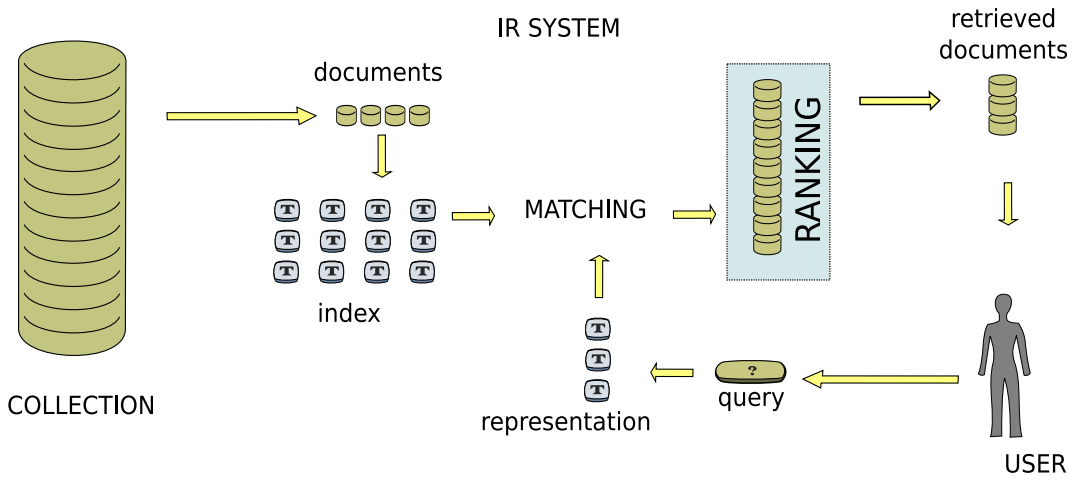

Figure 2.1: Scheme of the working of a basic IR system

1. The Query (definition 2.1).

2. Conditions required from a retrieved document. Both the task and the perception the user has about a document can be complex and partially unknown, so they are best split into different aspects for their study. Different aspects of relevance (like those described in subsection 2.2) can be taken into account for different tasks.

3. The language he or she is assumed to use. Lexical information from the collection is assumed to be an indication of the information the user would obtain from the text.

2.1.3 General Structure of an IR system

The basic problem that IR tries to solve involves two parts: a collection of information objects and a user with an information need. The solution to this problem is, of course, an *IR system* that mediates between the two, presenting the user with the relevant elements of the collection. The elements and structure of this system can vary from one approach to another, but are usually determined by a certain division of tasks.

Some of the key tasks that can be defined within the functioning of an IR system are the following:

Definition 2.4 (Indexing)

Indexing is the process of generating a suitable representation of documents, and organising

it for an efficient processing. This representation (the **index**) should be easier to manage than documents themselves, but have enough information about the subject they treat. The query also needs to be also represented in a similar way, to be compared (matched) with the documents.

Definition 2.5 (Matching)

Matching is the process of assessing a certain kind of relation between the representation of the query and the representation of each document. This relation is usually a query-document similarity, but, as we will show later (section 2.4) can be of other kind. The result of this is an amount that quantifies the degree of relevance (**score**) and allows to generate a list of the documents ordered according to their scores (**ranking**).

2.2 The Scope of This Work in the Context of IR

As mentioned in the last section, in this work we adopt a system-oriented approach, which means that we take a formulated query (instead of a user with an information need) as a starting point. It has been pointed out by Ingwersen [28] that the implication of a human user in the IR framework makes a drastic change of approach necessary: from one akin to "hard sciences" to another one that is closer to humanities. However, there is still an important part of IR dealing with computer systems and coded information. The restriction of the problem allows us to perform a deeper exploration of the concepts and problems involved.

As one of the consequences of choosing a system-oriented approach, we will not be concerned with some of the contextual aspects of **relevance** for a given task, but instead will adopt a concept that is not centred on the user but on the document itself: **aboutness**. In the next two subsections we define the two concepts.

Relevance

According to Saracevic in [29], relevance is a concept that, in spite of its ubiquity, has a rather tacit meaning, and a large number of different definitions were produced when a formalisation was attempted. Saracevic puts all these definition together in two ways: a minimal, abstract definition, and a maximal, extensive definition:

Definition 2.6 (Relevance (minimal))

Minimal definition of Relevance: Relevance is a measure of relatedness: it is the strength with which a set of objects P are related with a set of objects Q.

Definition 2.7 (Relevance (maximal))

Maximal definition of Relevance: Relevance is the $\{\mathbf{A}\}$ of $\{\mathbf{B}\}$ existing between $\{\mathbf{C}\}$ and $\{\mathbf{D}\}$ as determined by $\{\mathbf{E}\}$

where:

- $\{\mathbf{A}\} = \{$ measure, degree, estimate, $\ldots\}$
- $\{\mathbf{B}\} = \{$ correspondence, utility, fit, $\ldots\}$
- $\{\mathbf{C}\} = \{$ document, information provided, fact, $\ldots\}$
- $\{\mathbf{D}\} = \{$ query, request, information requirement, $\ldots\}$
- $\{\mathbf{E}\} = \{$ user, judge, information specialist, $\ldots\}$

Other aspects of relevance, like the genre of a document, its aesthetic characteristics, etc. can be used to enhance IR [30], but are closely related with the context of the search and user, and are therefore better accounted by a user-oriented approach than by a system-oriented one.

The definition of relevance, however concrete or abstract, necessarily includes a human user and its context. Only some aspects of relevance can be explicitly used in a system-oriented approach; namely those grouped by Mizzaro as one of the dimensions of relevance [31]. A system-oriented approach faces classification and categorisation problems similar to those defining library science, and can therefore adopt a concept borrowed from that field: *aboutness*

Aboutness

Aboutness was first defined in the realm of Information and Library Science, and has been closely related to the task of summarisation: obtaining and representing the topic treated in a document. Hutchins, in 1977, describes aboutness as a property of the documents including only the basic and general semantic contents [32]:

> To summarise the essential features of this approach to document 'aboutness', we suggest that for the purposes of information systems a summary of the total semantic content of a document is not what is needed. The primary aim of indexing is to provide readers with points of contact, leading them from what they know to what they wish to learn. In document analysis the most important parts of a document's semantic network are those elements that form the knowledge base upon which the writer builds the 'new' information he tends to convey.

Hutchins distinguishes two granularities in the use of particular terms: a "micro" usage that has to do with phrases or sentences, and a "macro" use having to do with a whole document; he claims that it is this *macro* semantic level, which he calls "theme", what defines what the document is about. To catch up with the current terminology, we will call it *topic*, and define it as follows.

Definition 2.8 (Topic)

A **topic** is a certain set of semantic relations that restrict the meaning of terms (or other information objects[2]). These relations are defined at a document, or set-of-documents level of granularity, as oposed to a micro level.

This will allow us to define aboutness as follows:

Definition 2.9 (Aboutness)

aboutness is the degree to which the elements (e.g. terms) of a document are used according to the restrictions imposed by a topic. This degree can also be defined as binary, so that a document can be simply said to be about something or not. This relation will be represented as:

$$Document \ \square\!\!\rightsquigarrow \ Topic \tag{2.1}$$

This definition complies with the formal characteristics that are required by formal accounts of IR ([33], [34]). Bruza *et al.* suggested to consider both documents and topics as sets of infons[3], so that aboutness becomes a binary relation with a number of convenient properties [36]:

[2] In this work we will be mainly concerned with *terms* as the minimum element, but the possibility remains open of applying the framework to other kind of elements

[3] **Infons** are the basic units of information, and contain an assertion like "a property holds for the elements of a set" [35]. The characteristics of infons that are relevant here, are that two things can be defined for them: a union operation $\cup$ and equivalence relation $\equiv$. They are defined for infons in a similar way as they are for sets.

1. Reflexivity

$$A \mathrel{\square\!\!\rightsquigarrow} A \tag{2.2}$$

2. Transitivity

$$[A \mathrel{\square\!\!\rightsquigarrow} B] \wedge [B \mathrel{\square\!\!\rightsquigarrow} C] \Rightarrow [A \mathrel{\square\!\!\rightsquigarrow} C] \tag{2.3}$$

3. Set Equivalence (for a given equivalence relation $\equiv$)

$$[A \mathrel{\square\!\!\rightsquigarrow} B] \wedge [B \equiv C] \Rightarrow [A \mathrel{\square\!\!\rightsquigarrow} C] \qquad [A \mathrel{\square\!\!\rightsquigarrow} B] \wedge [A \equiv C] \Rightarrow [C \mathrel{\square\!\!\rightsquigarrow} B] \tag{2.4}$$

4. Left Monotonic Union (for a given operation of union of sets of infons $\cup$)

$$[A \mathrel{\square\!\!\rightsquigarrow} B] \Rightarrow [A \cup C \mathrel{\square\!\!\rightsquigarrow} B] \tag{2.5}$$

5. Cut

$$[A \cup B \mathrel{\square\!\!\rightsquigarrow} C] \wedge [A \mathrel{\square\!\!\rightsquigarrow} B] \Rightarrow [A \mathrel{\square\!\!\rightsquigarrow} C] \tag{2.6}$$

These properties can be used to check an aboutness assessment technique. They will be used with that defined in chapter 5, as a sanity check for the method.

2.3 Use of Lexical Information in IR

Every methodology in textual IR has to do with lexical measurements: criteria for aboutness or relevance deal almost solely with the occurrence of particular terms in the text and the query. A novel scheme of lexical measurement can then be related to all existing methodologies, and should also suggest improvements to some of them. To give a brief outlook on how lexical information is used in IR, we will classify the type of approaches by the way each quantifies this information.

To illustrate how different methodologies could benefit from a principled approach to lexical measurements, let us present a rough classification of common IR methodologies. In figure 2.2 we show a succinct map with different kinds of IR models, where the method proposed in this thesis is situated.

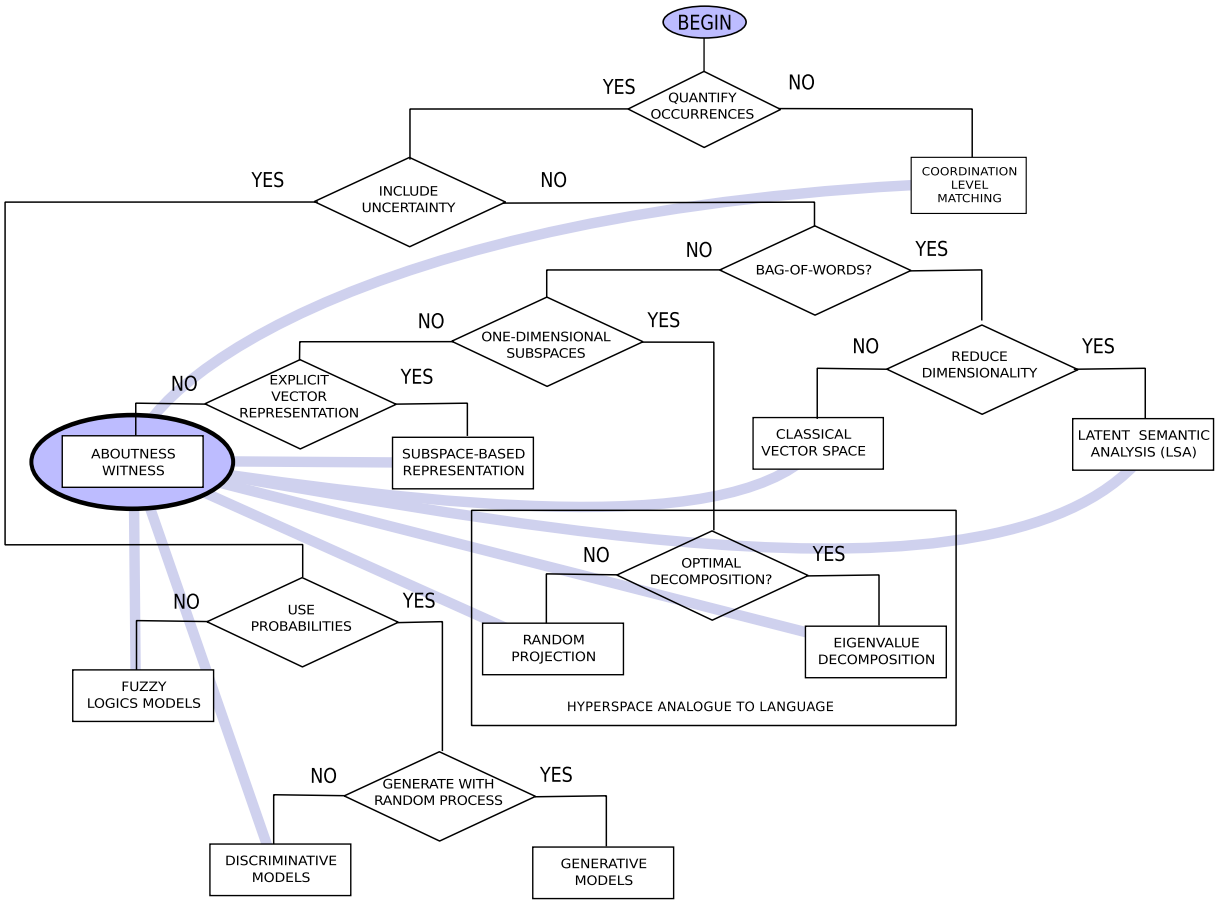

Figure 2.2: Overview of different approaches to IR. The one proposed in this work is labeled "Quantum-Like Lexical Measurements", and the soft lines to the others mean that they can be considered as a particular version of this approach

2.3.1 Coordination Level Matching

There are a few methods for textual IR that do not use probabilistic considerations. The first classifying criterion shown in figure 2.2 is whether the number of occurrences of a term is used (quantify occurrences), or just its presence. Methodologies using just presence or absence are referred to as **boolean**, and the most successful amongst those is *coordination level matching*. Queries are taken as boolean expressions referring to the presence or absence of keywords, which in turn define sets of documents fulfilling them (documents with the terms, documents without the terms). Using the distributive law between intersection and union, the query is translated into the intersection of several expressions, so documents can belong to the sets defined by a number of them, and are then ranked according to how many of this sets they belong to. This method is one of the earliest in IR, and was described in 1957 by Luhn [37], together with statistical improvements

which inspired many posterior models. It is worth noting the role of propositions referring to term usage in the document as the raw material for logic expressions, which is something we will also consider in this work.

2.3.2 Purely Geometrical Methods

Some of the methods used in IR do not deal directly with uncertainty, but use similarity and/or distance between representations instead. They all can be traced historically to the Vector Space Model:

Vector Space Models (VSM)

In 1975 Salton proposed one of the most influential models for IR: the Vector Space Model [16]. The basic idea of this methodology is that documents can be represented as vectors in a term space, so that a similarity function can be defined that measures how similar the term occurrences distribution are between two documents. The simplest representation for a document would be:

$$|D\rangle = \sum_i f(N_i \; in \; D)|t_i\rangle \tag{2.7}$$

where $|D\rangle$ is a vector representing the document (see appendix A for an explanation of this way of denoting vectors, called *Bra-Ket Notation* or *Dirac notation*) $f(N_i \; in \; D)$ is a real (usually non-negative) number; a function of the number of occurrences of term t_i in document D, and $\{|t_i\rangle\}$ is a basis spanning a term space. In terms of vector spaces, this means that the space of terms spanned by $\{|t_i\rangle\}$ is *dual* to the space of documents: each term can be regarded as a functional that linearly assigns a real number to every document. This number can be interpreted as a contribution to the term-vector to the representation of the document-vector. Similarities between documents can be obtained from these numbers $f(N_i \; inD)$, usually used as the L^1 norm[4]

Some of the weighting functions that define a vector representation for a document can be seen in table 2.1. With a sensible weight assignment scheme, this model gives good retrieval results,

[4] L^1 norm consists simply on the sum of the entries of a vector [38].

Scheme	Formula	Reference
TF	N_i/L_D	[16]
BM25	$(K_1+1)N_i/(N_iK_1+b+(1-b)\frac{L_avg}{L_D})$	[39]
TF·IDF	$(N_i/L_D)\cdot\log(N_D/N_{D\ with\ t_i})$	[40]
Pivoted Normalisation	$\frac{N_i\cdot\log(N_D/N_{D\ with\ t_i})}{((1-slope)\cdot pivot+pivot*L_D)}$	[41]

Table 2.1: Functions of the frequency of occurrence used to define vector-space representations of documents. N_i is the number of occurrences of the term, L_D is the length of the document, and $N_{D\ with\ t}$ is the number of documents where term t is present. K_1 is a free parameter characteristic of the Okapi method, L_avg is the average length of the documents in the collection. $slope$ and $pivot$ are also free parameters that need to be fitted in the pivoted normalisation method.

and the choice of these weighting schemes offers the possibility of taking it to the realm of statistical models ([42], [43], [44]). Finding an appropriate weighting scheme for terms, however, is a difficult problem, as is discussed in [45].

Kernel Models Based on Dimensionality Reduction

As a refinement of vector space models, schemes have been proposed to use the information of term usage in a training set (that could consist of the whole collection) to get more information from term occurrences. In the following sections we will show the basis on which some of these models are built, both for the case of bag-of-words approaches like the classical *Vector Space Model* (VSM) and *Latent Semantic Indexing*, and for co-occurrence models, like *Hyperspace Analogue to Language* and novel quantum-inspired approaches like *Spanned Subspace Representations* and *Aboutness Witness* (this last one being proposed in this thesis).

In the original VSM term occurrences were assumed to be independent and geometrically equivalent features to represent documents. This can be seen as an arbitrary deformation of the space: the presence of two synonyms should tell less than the presence of two unrelated terms; the presence of a noun brings more information than that of a preposition. There is a simple way to *deform* the space where documents are represented to fit the particularities of the features used. The operator that performs this deformation of the term space is called **Kernel**. In the Generalised Vector Space model, a positive-definite kernel matrix weighting can be introduced in a similarity formula, to

account for term weighting and relations between terms [46]:

$$S(D_1, D_2) = \frac{\sum_{i,j} K_{i,j} \langle D_1|t_i\rangle\langle t_j|D_2\rangle}{\sqrt{\left(\sum_{i,j} K_{i,j}\langle D_1|t_i\rangle\langle t_j|D_1\rangle\right)\left(\sum_{i,j} K_{i,j}\langle D_2|t_i\rangle\langle t_j|D_2\rangle\right)}} \tag{2.8}$$

Where hermitian[5] matrix $K_{i,j}$ carries information about the importance of terms and the redundancy between them. Vectors $|D_1\rangle$ and $|t_i\rangle$ are representation of documents and terms in their dual spaces (the differenes between representing them as a bra $\langle\cdot|$ or a ket $|\cdot\rangle$ is not important here; it only becomes essential when complex vectors are used.

When the entries of a kernel $K_{i,j}$ depend only on the sub-indices i and j, using it is equivalent to expressing terms themselves as weighted combinations of a latent independent and unbiased basis:

$$|t_i\rangle = \sum_j L_{i,j}|\lambda_j\rangle = \sum_j |\lambda_j\rangle\langle\lambda_j|t_i\rangle \tag{2.9}$$

where kernel K is a quadratic form of the latent coefficients, plus some latent weights:

$$K_{i,j} = \sum_k L_{i,k} L_{j,k} w_k \tag{2.10}$$

which would made the similarity function:

$$S(D_1, D_2) = \frac{\sum_{i,j,k}\langle D_1|t_i\rangle\langle t_i|\lambda_k\rangle w_k \langle\lambda_k|t_j\rangle\langle t_j|D_2\rangle}{\sqrt{\left(\sum_{i,j,k}\langle D_1|t_i\rangle\langle t_i|\lambda_k\rangle w_k\langle\lambda_k|t_j\rangle\langle t_j|D_1\rangle\right)\left(\sum_{i,j}\langle D_1|t_i\rangle\langle t_i|\lambda_k\rangle w_k\langle\lambda_k|t_j\rangle\langle t_j|D_2\rangle\right)}} \tag{2.11}$$

Since VSM relies on an inner product to define similarity for a matching scheme, different methodologies can be defined depending on the function defining the inner product, namely the kernel. Kernels are defined by an optimal matrix $K_{i,j}$ so that the product is defined as:

$$a \bullet_K b = \sum_{i,j} a_i K_{i,j} b_j \tag{2.12}$$

The matrix K is optimal in the sense that its rank is less than the number of terms, while inner products computed with it are still as similar as possible as computed with a simple entry-by-entry

[5] A hermitian matrix is one that is identical to its conjugate transpose. It is a generalisation of symmetric real matrices to complex numbers

inner product (noted simply as a product $d_i \bullet d_j$). Similarity is here represented as a similarity function between matrices $\Delta(x, y)$ with a real value that is maximum with $x_{i,j} = y_{i,j}$ and diminishes according to how different x and y are:

$$\forall K' \neq K, \Delta(d_i \bullet_K d_j, d_i \bullet d_j) > \Delta(d_i \bullet_{K'} d_j, d_i \bullet d_j) \tag{2.13}$$

Different functions, with different constrains for K produce a range of methods:

1. Latent Semantic Indexing (LSI) Consists in representing documents with vectors defined in an optimal basis set. This basis set is computed to reproduce optimally a matrix of term-term cosine similarities between terms, while limiting the size of the basis set. The kernel is in this case a projector on an optimal subspace [47].

2. Non-Negative Matrix Factorisation (NMF): Defines a non-orthogonal basis for which all the coefficients of the representation of single term will be non-negative. The basis are also optimal to reproduce the matrix of term-term cosine similarities under the restriction of non-negativity [48].

3. Probabilistic Latent Semantic Indexing (PLSI): A set of non-orthogonal basis are also defined to keep non-negativity of the term coefficients, like in the case of NMF, only this method aims to reproduce mutual information between occurrences of different terms, instead of cosine similarities [49]. This method can also be considered in the category of generative probabilistic models.

Hyperspace Analogue to Language

An attempt to use co-occurrence of terms in the text at a fine-grained level was made by Lund and Burgess (see [50]). This approach is called Hyperspace Analogue to language (HAL). It consists in sliding a window of fixed width through all the documents in a collection, and count co-occurrences of all terms within this window. The result is a huge matrix of co-occurrence between all the terms, which can be used to generate a term-term kernel. Dimensionality reduction techniques such as eigenvalue decomposition [50] and random projection [51] have been used to remove noise and produce a compact representation of the kernels.

Aboutness Witness

One of the methods suggested by this work (developed in chapter 5) will be based on a scheme to measure both occurrence and co-occurrence of terms. Operators representing this measurement (called Selective Erasers) will be combined to define a witness operator that will assign real numbers (scores) to documents. This method will implicitly use co-occurrence information gathered from a collection of corpus, which amounts to include a co-occurrence kernel. This method does not require an explicit geometrical representation of either documents nor terms, but can support it; it can also be related to logics through a suitable definition of uncertain conditional (a concept that will be explained in section 2.4), and can also be given a probabilistic iterpretation, as well as include elements from probabilistic techniques.

Subspace Models

Traditional VSM can be said to represent documents and queries with one-dimensional subspaces within a total Hilbert Space[6] However, QT suggests that higher dimensional subspaces can be also used to represent objects, and this actually bring other desirable features to the representation. Amongst the models following this idea, Mellucci proposed one where queries are represented as a high-dimensional subspace while documents are one-dimensional [52], while Piwowarski and Lalmas propose a scheme that allows high-dimensional representation of both [53].

2.3.3 Methods Involving Uncertainty

The first classifying question we can make about statistical models, is whether the statistics of the whole collection are considered or not.

Fuzzy Logics Models

One way of not using statistics at all is with fuzzy logics models. They consist in using non-binary term weights as degrees of pertinence to fuzzy sets, and apply the corresponding generalisation

[6] a **Hilbert Space** is a linear, decomposable vector space where an inner product is defined

scheme to extend classical boolean methods [54]. Weights can be seen in fuzzy logics as valuations of propositions replacing binary "false or true" with an ordered set (for example, numbers between 0 and 1). The method can be applied by defining an appropriate mapping from Boolean operations to this set of valuations, for example:

$$w(P \wedge Q) = \min(w(P), w(Q)) \qquad w(P \vee Q) = \max(w(P), w(Q)) \tag{2.14}$$

where $w(P)$ is the continuous valuation (in an interval from 0 to 1) of a proposition P. In section 3.3.2 we will discuss valuations, and how a more general kind of valuations can introduce aspects of quantum logics to this kind of models.

Probabilistic Boolean Retrieval

Not using the statistics of the whole collection, on the other hand, does not exclude the possibility of using probabilistic reasoning to extend Boolean methods. In 1976, Robertson and Spärck-Jones developed a simple probabilistic model usually called Binary Independence Retrieval (BIR) model [19] which can be seen as a version of VSM that is not built on a notion of distance, but on probabilistic grounds. In the simplest version of this model, a weight for terms is computed as the logarithm of the odds for the presence in a relevant document, against that in a non-relevant document:

$$w(t) = \log\left(\frac{P(t\ in\ D|L)}{P(t\ in\ D|\neg L)}\right) \approx \log\left(\frac{\frac{r}{R-r}}{\frac{n-r}{N-n+R-r}}\right) \tag{2.15}$$

where t is the term, D is the document, L is the event of a user marking the document as relevant. r is the number of relevant documents with term t. R the number of total relevant documents, n the number of documents with term t, and N the total number of documents. An assumption is made of *linked dependence* [55] between occurrences of different terms, meaning that the log-likelihood of a document being relevant can simply be obtained as the sum of the weights of the terms in it:

$$\log\left(\frac{P(D|L)}{P(D|\neg L)}\right) \approx \sum_{t\ in\ D} \log\left(\frac{P(t\ in\ D|L)}{P(t\ in\ D|\neg L)}\right) \tag{2.16}$$

Discriminative Models

Usage of terms in a whole collection gives valuable information about how the occurrences of a term can be a clue of relevance. A simple way of using this information is assuming a sensible form for the distributions of the occurrences of the term in both the set of relevant documents (elite) and in the rest of the collection. Then the likelihood can be computed that a particular sampling corresponds to the elite term distribution. [56] This amounts to compute a conditional probability: given the sampling given by the document, what is the odds that the document is relevant instead of non-relevant? This question gives rise to *discriminative models*. The first model in this class was the *two Poisson model* proposed by Harter in 1975 [43], which assumes a Poisson distribution for an elite set of relevant documents, and another Poisson distribution for the others. Other discriminative models use multinomial distributions to approximate conditional probabilities instead [57].

Generative Models

From the seminal work of Shannon on language as a random source of information [58], several accounts of linguistics have tried to use a model of natural language text as generated by a random process. This was an important influence in computational linguistics, and arrived into IR with Ponte and Croft Language Model for IR [59] where a Language Model for IR is devised as non-parametric models where the statistical regularities of the generating process are learned from text itself: a probabilistic model is learned from the document, and the probability that the query is generated with this model is computed, as an approximation to the probability of relevance. This initial generative model did not include the notion of topics, but this has been introduced int newer methods such as Latent Dirichlet Allocation [60], where a continuum of topics is defined, so that a density of topics is found for a given document, and Probabilistic Latent Semantic Analysis [49], where optimal topics are defined by a maximum likelihood criterion.

2.4 Uncertain Conditional as a Unifying Approach

All the above mentioned methods point at computing the probability (or a monotonic function of it, like likelihood) that the document is about a topic defined by a query. A topic would be represented by a characteristic probability distribution of terms, so it can be recognised statistically or otherwise. This definition has proved to be very useful, but is remarkably shallow in terms of semantics and logics.. In 1976 van Rijsbergen proposed a much better starting point: casting aboutness in terms of a certain logical conditional: **if a document implies the query, then it is *about* underlying topic**.

$$(D \boxdot\!\!\rightsquigarrow Q) \iff (\mathbf{R}(D) \rightarrow \mathbf{R}(Q)) \tag{2.17}$$

where $\boxdot\!\!\rightsquigarrow$ is the aboutness relation and $\mathbf{R}(\cdot)$ is a representation in terms of objects in which implication $\rightarrow$ relations are defined. Since the necessity of a non-crisp quantification of aboutness has been both experimentally and theoretically found, van Rijsbergen also proposed that it should be possible to put this relation in terms of a conditional probability, such that:

$$\mathrm{P}(D \boxdot\!\!\rightsquigarrow Q) = \mathrm{P}(\mathbf{R}(Q)|\mathbf{R}(D)) \tag{2.18}$$

This statement restricts enormously what the formal representation of aboutness is, but still does not specify it completely, because what *imply* would mean is not yet specified. This approach was the root of a wealth in logic-oriented IR research, and can be also related to the other kinds of methods: geometric (vector spaces, via projector-oriented logic) [61] and probabilistic (via inference) [62].

2.4.1 Logics, Semantics and Implication

Logic deals with propositions, and most importantly, their relations; while semantics deals with the relation between those propositions and the subject they are dealing; with something external they are representing (their meaning). Since a logic is an extremely abstract construct, it can be built on rules alone, without any necessary relation with a useful meaning.

2.4.2 Information in the Document vs. Information in the Query

In this work we will stick to Dretske's definition of information [63]. Dretske tries to represent knowledge as the flow of information, and claims that for this ends, it is convenient to define information as an objective commodity, something whose existence is independent of the interpretative activities of conscious agents. His definition is:

Definition 2.10 (Information)

A signal r carries the **information** that $a \boxed{R} b$ when

$$P(a \boxed{R} b|r) = 1 \tag{2.19}$$

where $\boxed{R}$ is a relation.

We could put the problem of aboutness in terms of presence of information; we can say that a document is about a topic when it carrries information about particular relations between some of the concepts that define the topic, and also between these concepts and others that do not belong to the topic.

Observe that what was defined above is information *about* a proposition ($a \boxed{R} b$). It has to be noted that the quantification of information is a different problem altogether, and is not related to particular messages, but is statistical in nature.

Shannon and Weaver's definition involves an average on the possible transmitted messages [58]. In their theory of communication, the measure of information "produced" in a possible measurement process should be proportional to minus the logarithm of the shrinkage ratio of the set of possible states, and should be averaged on the possible measurements. Since a probability can be thought itself as the inverse of a shrinkage ratio of the number of possibilities, the measure proposed for a set of mutually excluding messages would be:

$$P_i = \frac{|\{\text{possibilities after choosing i}\}|}{|\{\text{all possibilities}\}|} \qquad H(\{P_i\}) = \sum_i -P_i \log(P_i) \tag{2.20}$$

Quantity H is usually called **Shannon Entropy** (or simply *entropy*) and is actually a decreasing measure of the amount of information; it decreases as certainty increases. A measure of the amount

of information would be best described as

$$I(\{P_i\}) = H(\text{uniform}) - H(\{P_i\}) \tag{2.21}$$

When a document satisfies an information need, it should provide some information that was not present when the query was formulated. So far, Shannon account is enough. But the information should be also *about* the topic at hand; this requires to go further and use Dretske's definition. The fact, for example, that a retrieved text is presented in one font or another is indeed information, but is not information about the topic. Note that this is closely related to Hutchins remark mentioned in 2.2, picked in this work as one of the arguments in favour of the use of the concept of aboutness.

We could think of both query and document as descriptions that define sets of possible *situations*[7], or, better, adequately *describe* situations from such set. In this way, a definition of relevant documents consistent with relation (2.17) can be that **a relevant document describes only situations that are also described by the query** (although it describes them in more detail).

The query has little information, so it could be said that describes situations very partially, and the set of situations that are adequately described by the query is large. It is reasonable to think that the user has some idea of the situations in this set when he formulates the query. Then, he gets a document, and the information in it makes him discard some of the situations he had in mind, thus increasing his amount of information. However, the document could also make him think of situations he did not think when he formulate the query first. This case is, in terms of Shannon information, a bit problematic, as is shown in figure 2.3.

The problem of managing partial relevance, a central one in IR, can be tacked by using a different kind of conditional, as we will show in the next subsection.

[7] A *situation*, is a complex of entities and relations that restricts only incompletely a state of affairs, or possible world (or, equivalently, is compatible to several of them). The reader can, however, find a definition more oriented to IR in [64]

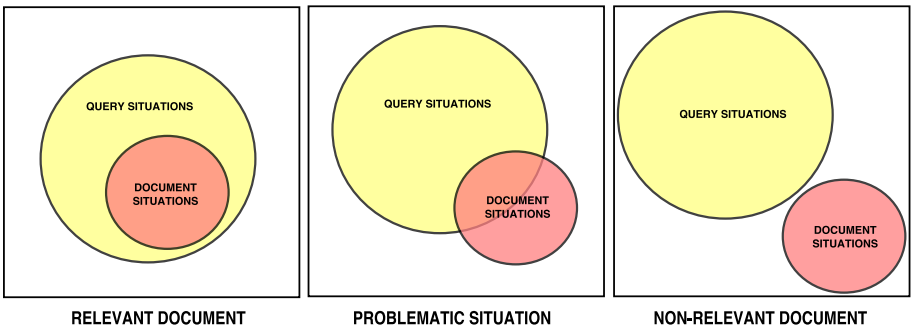

Figure 2.3: Restriction of the set of possible situations consistent adequately describable with the knowledge of the user

The case in the middle appears intuitively as a partially relevant document, but is something problematic in terms of information, because the document bring into consideration states that were not initially considered, and usual measurement of information deals only with discarding of possibitlities.

2.4.3 Evaluating an Uncertain Conditional

Intuitively, we could say that reading a document, the user could revise its initial idea of the possible situations, and consider, in the light of the new information, a wider *initial* set of possible situations, to contrast with the final set. This amounts to perform what is called the *Ramsey Test*. This is defined in [65, page 29]

Definition 2.11 (Ramsey Test)

Ramsey Test: To evaluate $[A \rightarrow B]$

1. Take my present system of beliefs, and add to it so as to make $\mathrm{P}(A) = 1$.
2. Allow this addition to influence the rest of the system in the most natural, conservative manner
3. See whether doing so results in a higher probability $\mathrm{P}(B)$

In the case of the query and the semi-relevant document, a similar procedure was proposed by van Rijsbergen in [66] for quantifying relevance: enlarge the set of possible situations described by the query in a natural, conservative way[8] until every situation described by the document is also in the enlarged set ($\mathrm{P}(s(D)|s(Q)) = 1$, where $s(x)$ means "situation s is partially described by x")

[8] just as Bennet in [65], we do not specify for the moment what the natural, conservative way of augmenting a set of situations is, because that is one of the outcomes we expect from this work

and check how much the set of possible situations had to be enlarged. The definition made by van Rijsbergen is:

Definition 2.12 (Logical Uncertainty Principle)

Logical Uncertainty Principle: Given any sentences x and y: a measure of the uncertainty of $y \rightarrow x$ relative to a given data set is determined by the minimal extent to which we have to add information to the data set, to establish the truth of $y \rightarrow x$ [67].

This principle is ilustrated in figure 2.4.

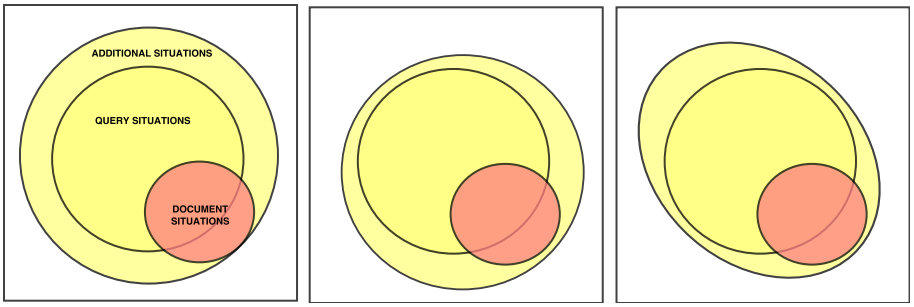

Figure 2.4: Using a Ramsey-like test to assess relevance $s(D) \rightarrow s(Q)$
The set of possible situations describable by the query is augmented until every situation describable by the document fits in. The three different figures illustrate that there could be several ways of doing it. The shape of the enclosing set of additional situations will depend on the geometry of the space of situations defined: the addition should not be arbitrary.

Numerically, this evaluation should behave like a conditional probability [66], so it would be compatible with the usual operational assumption that for a given topic there is a characteristic probability distribution of document important features (like keyword occurrences).

The Uncertain Conditional, in conclusion, may provide a unifying backbone for IR theory, but imposes nontrivial demand on the logical aspects of the theory. These conditions have been part of the motivation to turn to QT for useful analogies [10] and have been a source of inspiration for this work itself.

2.5 What QT has to offer to IR

QT provides a number of concepts that are potentially useful for a better description of concepts in areas outside physics. They are an extension, or generalisation, of existing concepts of probability

theory or even logic. In this section we are going to divide them according to which extension to the boolean or classical notions they are based on: Some are based in taking *probability amplitudes* as the basic magnitudes from which probabilities are defined; some are based on the non-local description of quantum systems which is particularly well-suited to describe correlation; and finally, some are based in the features of logic itself that are suggested by the quantum description of reality.

Most of this models are beyond the scope of this work, but are worth mentioning because they are also inspired in QT. The last category of models, however, which is referred to as "Relations between Lexical Measurements", is in fact the main contribution of this work. It is briefly described here, but will be fully developed in chapters 4 and 5.

2.5.1 From Probabilities to Amplitudes

One of the key features of QT is the impossibility of getting a complete information of the state of a physical system, a consequence of the intrinsic incompatibility of some observables. The usual example of this is the impossibility of determining the position and velocity of an elementary particle. Any measurement of the position would necessarily destroy the information about velocity, and *vice versa*. Representing a state of knowledge that is complete in some observables but uncertain in others led directly to the use of *probability amplitudes*. We define them as follows:

Definition 2.13 (Probability Amplitude)
A **Probability Amplitude** is a complex number with norm less or equal to 1. Its interpretation can be explained considering the probabilities of two independent possible outcomes of a common initial state, with the possibility of intermediate states between them. How they can be computed is dictated by *Feynman's Rules* [68]:

1. Probabilities are secondary quantities. They are squared norm of complex Amplitudes, which are the primary quantities.
2. When it is possible to tell the intermediate steps, you sum probabilities
3. When it is not possible to tell the intermediate steps, you sum amplitudes

Complex amplitudes allow the possibility to *save* (or *hide*) probabilistic unavailable information in the phase, which can manifest itself indirectly through *interference* when the third law of Feynman is applied. An extremelly interesting discussion of the necessity of doing so (together with a equally interesting alternative) was proposed by Wooters in [69]. When it is accepted that the probabilistic information about an incompatible observable *has* to be encoded in the representation, the minimum formal requirements clearly suggest to do it with complex numbers, as was shown by Goyal *et al.* [70].

This fact was the inspiration of a successful attempt to overcome the main limitation of PRP: document relevance independence. Zuccon *et al.* proposed in [71] a *Quantum Probability Ranking Principle* (QPRP) where amplitudes are assigned to the possibility of relevance of a document, which are computed according to the third law of Feynman, modelling the effect of previous judgements on a current one through interference.

Other attempts to use similar concepts for IR, without explicitly regarding them as related to QT. Park *et al.* suggest in [72] a method for using information about the position of terms that makes use of this. A Fourier transform of the sequence representation of the document provides an efficient and elegant way of representing information about positions of the occurrences of terms within the sequence of text. The square of the amplitude of the coefficients of such representation can be interpreted as weighted probabilities, and the phase bears information about a *frequency of occurrence*. Fourier transform is a particular case of Wavelet transforms, and other kinds of encoding can be performed with other wavelets, so this proposal offers a huge variety of methodologies to explore. Park's method is largely heuristic, but can be interpreted in terms of non-compatible observables and their quantum-like representation.

2.5.2 From Correlation to Entanglement

In 1935 Einstein, Podolsky and Rosen [73] presented what has been called the EPR paradox, claiming that QT was not a complete physical theory. What they meant, is that "*Every element of the physical reality does not have a counterpart in the theory*"; there were magnitudes whose value could not be *determined* in certain states of a physical system: not because they were not known, but because, according to QT, the magnitude itself did not make sense in the experimental setting

at hand.

In an elegant analysis of the particular *Gedankenexperiment* proposed by Einstein *et al.*, John S. Bell showed in [74] that the paradox went further than first thought: it implied that in some cases the only magnitudes that made sense were not abscribed to a certain point at a certain time, but were spread in different places, in a way he dubbed *non-local*. These non-local magnitudes, strangely enough, do not mess with causality, but do imply correlations beyond those describable by Newtonian physics, and even call for a certain kind of holistic description [75].

There has been in the last decades a clear interest in using this correlation-beyond-Newtonian to describe relations in cognitive sciences and linguistics. Some of the early works were on a quantum-like description of the decision process [76], a problem that has been developed further by Busemayer *et al.* in [77] and [78]. Cognitive states have also been described with quantum concepts in some works by Khrennikov *et al.* ([79], [80]) and

In an area that is closer to IR, some semantic effects have also been described as quantum-like. Peter Bruza *et al.* explored in [18] some relations between terms through the concept they called "spooky action-at-distance" in the human mental lexicon,

2.5.3 From Boolean Logic to Quantum Logic

A different approach to IR than those described up to this point, is to take it near to the domain of databases. Even though databases address a different problem, the need to bring them closer to the necessities of a human user has led researches to experiment with deviations of Boolean logic that are highly scalable and distributable, two conditions that are present in IR but acute in database praxis. This has inspired Schmidt *et al.* ([81] and [82]) to build a database query languages that include logical connectors that behave like those in quantum logics. In this work, the concepts from classical databases are mapped to quantum counterparts, obtaining in this way a natural and robust framework to use also probabilities that is missing in purely Boolean methods. A similar mapping could be used in diverse schemes of fuzzy logics, but it is shown that this brings about issues of robustness and error-proneness.

Relations between Lexical Measurements

In this category of Quantum-Logical models we can also situate those developed in this thesis: they are all based on relations between lexical measurements. Lexical measurements themselves can be defined as sharing key mathematical properties of quantum measurements, and can therefore be represented as such. The results of these measurements define propositions, and these propositions of the form *measurement X produced result Y* can be used as elements of a logic, together with relations and operations between them. The information contained in a document, according to definition 2.10, can be relations between lexical measurements, and used to define aboutness with (2.17).

2.6 Summary

In this chapter, we have very briefly outlined the state of the art of IR in the aspects that are relevant to the proposed approach. We stated the restrictions on the basic problem of IR that we have imposed to reach a level of simplicity that is treatable in depth. These involve mostly focusing on the system side of the IR problem, and dealing with aboutness instead of relevance. Then, we have shown some of the relevant features of current and past models of IR, showing how some of them can be generalised with concepts from QT, and how the use of concepts from QT suggests tthe development of new models as well, including those based on relations between lexical measurements developed in this thesis. Then, concept that brings some theoretical unity to all of these methods as also presented: the logical uncertainty principle, which plays be an important role in the basis of this work. Finally, the basic aspects in which concepts from QT can be used to extend existing models are briefly described, including the logical motivation of those proposed in this thesis.

In the next chapter, a very general theoretical view is presented, that will revisit some of the concepts discussed so far, and will point at the way they could be used to build a new theoretical account of lexical measurements for IR.

Chapter 3

Measurements and Information Retrieval

The approach to lexical measurements followed in this thesis is an *ab initio* approach: it starts from first principles. Measurement is taken as the fundamental concept on which the rest is built, and measurement itself is studied from a general and abstract mathematical (and logical) perspective. In this chapter, the notion of measurement is presented formally and in detail, with an emphasis on characteristics and conditions that are important for a lexical description of text. In particular, the importance of order relations and how measurement is related to logics will be discussed. These characteristics and conditions will be the basis on which we will build a scheme for lexical measurement in chapter 4.

In section 3.1, the motivation for the choice of measurement as the fundamental concept of this work is explained. In section 3.2, we explain the basic ideas about measurement from a formal point of view, which will be developed further for section 3.3. In section 3.4 we show how the existence of incompatible measurements can be accounted in terms of logics, and what consequences it has on descriptions of measurement events. Finaly, in section 3.5 we outline how all these concepts can be applied to Information Retrieval, and how they can provide a new starting point for the development of both theory and practical applications. These outlines will be developed in the following chapters. A concluding summary of the concepts introduced in this chapter can be found in 3.6.

3.1 Why Measurement?

When a person takes a newspaper without the intention to read it thoroughly, she or he will take a quick and shallow look at it, so anything interesting, recognised as some picture or word, can call for the attention of the reader. This way of reading has been called *scanning* [83], and is thought to be very usual when reading web pages [84]. In the early stages of library science, retrieval schemes were intensive in human work and could not rely on powerful processing resources, so a thorough reading of the documents by a librarian was necessary to produce a highly informative index. The selection of keywords proved to be a task that was difficult to automate. However, the advent of full-text search [85] brought about more machine-intensive methodologies that do not require a human annotator anymore. A machine is probably faster than a human to process large amounts of text, but is in most cases less precise at extracting meaning from it. When a human is presented a text in a language she or he understands, every word will appear to him as potentially charged with meaning since the first stages of perception, while in most automatic text processing schemes, any assignation of meaning occurs very late in the process.

The proposal of a new approach to lexical measurement as starting point of an IR system, is mainly motivated by the following assumption:

Assumption 1

The occurrence of a particular word in a document, when meaningful for a reader with a certain interest, will call his attention to the surrounding text. A lexical measurement can be defined to reproduce this phenomenon. Making an automatic system work in this way will enhance its ability to capture evidence about what the topic the text can be relevant to.

The fact that in an IR context the user is expected to read a potentially relevant document with a topic in mind [86], makes it reasonable to assume that a representation should reflect the existence of different topical points of view. These points of view might, furthermore, be incompatible to each other. Intuitively, it is this incompatibility that suggests adopting a quantum analogy: for a given physical system, QT allows the description of multiple sets of measurements that are internally compatible, but incompatible[1] to each other [61].

[1] *incompatible* measurements here means measurement that interfere introducing uncertainty in each other's outcome. This will be defined thoroughly in section 3.4

3.2 The General Problem of Measurement

Measurement is, as Kelvin pointed out in [87] (cited in [88]) a fundamental and key point of any scientific activity. However, its fundamental character is rarely discussed, except perhaps in the case of Quantum Theory, where its problematic nature makes discussion necessary. A very loose definition of measurement would be the assignation of a certain value to a characteristic (observable) of a considered system.

According to Carnap in [89], there are three kinds of scientific concepts: **classificatory, comparative** and **quantitative**. Classificatory concepts are exclusively descriptive characteristics of objects or states of objects that allow to compare them in a qualitative manner as similar or different only. Comparative concepts introduce also order relations between characteristics, so that object X could be *more* α than object Z, where α is a comparative characteristic (for example, X can be heavier, harder, bigger, etc. than Z). Quantitative characteristics are endorsed with a richer structure, where addition and subtraction are defined, and differences, for example, can also be compared. Measurement is usually related only to this third kind of concepts, even though sometimes it is more useful to define it as comparative or classificatory.

Norbert Robert Campbell, one of the first and most important theoreticians of measurement [90] seems to restrict his definition of measurement to the third kind of scientific concepts: quantitative [91]. According to Campbell, a measurement is *the assignation of a numeric value to a certain observable*, where mathematical relations within the set of numbers correspond to empirical relations within the set of values of the observable. He states three laws of measurement, and three corresponding rules that must be followed:

1. **Law: Ordering** Measurement involves order relations **Rule:** Order relations between magnitudes correspond to order relations of the numeric values assigned.

   What is referred to as *order relations* is a transitive, anti-symmetric relation such as “bigger-smaller than”, “is ancestor-descendant of ”, "predecessor-successor of", etc.

2. **Law: Additivity** Measurement involves additive quantities. **Rule:** When a standard value for an additive magnitude is chosen, a scale can be built by putting together independent systems with the same value of the (additive) magnitude. In Figure 3.1 a scale made

with systems of equal length is shown, and how this scale is used to establish the length of an object within that scale.

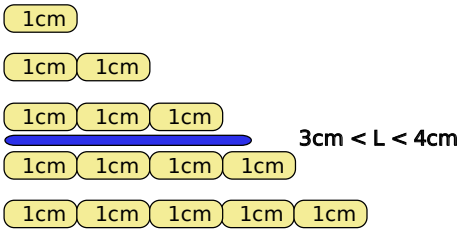

Figure 3.1: A scale for length made with standard systems of $1cm$. The length (L) of an unknown object is found to lie between $3cm$ and $4cm$; it is usually said to be simply $3cm$, and is understood that it lies between this value and the next (here no rounding is done, since it is not known whether it is closer to $3cm$ or to $4cm$)

3. **Law: Multiple Scales** For every standard chosen to build a scale, a sub-standard can be found that allows to build a finer scale that includes the values on the initial scale.

   This third law does not, as Campbell notes, necessarily hold for every measurement. It implies that the scale can be split indefinitely in finer and finer scales, just as the set of rational numbers. This law should hold, then, for magnitudes that can vary continuously. In Campbell's book [91] there is no mention of a *third rule* of measurement. However, the discussion about changing standards through multiplication in page 54 can be considered as the corresponding *third rule of measurement*: that a change from standard A to standard B would be, operationally, carried through the choice of a *sub-standard* X that includes the scales of both A and B. This is explained in example 1.

Example 1

If standard A is $1cm$ and standard B is $1in$, sub-standard X can be chosen to be a common divisor of these quantities. It should fulfill the following relations

$$N_1 X \leqslant 1cm \leqslant (N_1 + 1)X \tag{3.1}$$

$$N_2 X \leqslant 1in \leqslant (N_2 + 1)X \tag{3.2}$$

where N_1 and N_2 are natural numbers. When a length is found to be, for example, be-

tween $3cm$ and $4cm$ (like that in picture 3.1), this would mean:

$$3N_1X \leqslant L \leqslant 4(N_1+1)X$$
$$\Rightarrow (3N_1in \leqslant (N_2+1)L) \wedge (N_2L \leqslant 4(N_1+1)in) \quad (3.3)$$

Note that this relation involves only natural numbers, but if L is not to be measured against standard $1in$ but computed in inches from the measurement $1cm$, a division must be performed:

$$3\frac{N_1}{N_2+1}in \leqslant L \leqslant 4\frac{N_1+1}{N_2}in \quad (3.4)$$

As a smaller sub-standard X is chosen, numbers N_1 and N_2 will become bigger, and the range where L is located will approach a minimum width asymptotically.

3.2.1 Logical Aspects of Measurement: Subsumption Relations between Results

Campbell's third law (3) refers to a relation between scales: *some scales include others*. This relation can be used to define subsumption relations; this can be done by considering measurements as partitions of the set of all possible measurements into subsets.

In the case depicted in Figure 3.1, two of the elements of the scale ($3cm$ and $4cm$) define the following sets:

$$S_{3cm+} = \{l | l \geqslant 3cm\} \quad (3.5)$$
$$S_{4cm-} = \{l | l \leqslant 4cm\} \quad (3.6)$$

These two sets are clearly overlapping, and their conjunction is precisely the set where L can be found according to the measurement. A proposition $P(L, cm)$ can be defined, referring to length L and the standard cm:

$$P(L, cm) = (L \in S_{3cm+} \cap S_{4cm-}) \quad (3.7)$$

If a smaller standard is chosen to measure against, then a different relation is found for L, for example $33mm \leqslant L \leqslant 34mm$. The set obtained with this sub-standard mm would be included

in the set obtained by the initial standard cm, and we can say that this measurement *implies* the former. A proposition similar to $P(L, cm)$ can be stated about standard mm:

$$P(L, mm) = (L \in S_{33mm+} \cap S_{34mm-}) \tag{3.8}$$

It is easy to show that the subset defined with the standard cm includes that defined with substandard mm, so one of the propositions subsumes the other. Formally :

$$((S_{33mm+} \cap S_{34mm-}) \subset (S_{3cm+} \cap S_{4cm-})) \iff (P(L, mm) \supset P(L, cm)) \tag{3.9}$$

Where we use the same symbol $\supset$ that is used as an inclusion relation between sets; here it symbolises *subsumption* between propositions, or, as is also called in logics, material implication [65, page 20].

The partition of the set of possible results (in the example, all the possible lengths) in subsets defines a partially ordered structure, since inclusion is an order relation. Adding an empty set, and a set that contains every possible result, a structure called a lattice [92] is obtained. And since it is possible to define a subsumption relation within this lattice that we can regard as an implication, this is also a **logic**. Such sets-based logic is, in fact, a Boolean logic [61]. Figure 3.2 shows an example of such lattice.

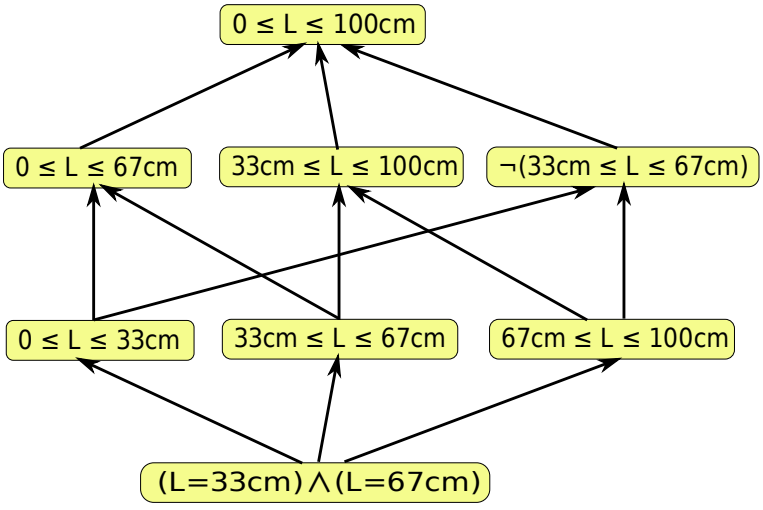

Figure 3.2: Boolean Lattice made with propositions involving three elementary length measurements. Arrows $\rightarrow$ represent inclusion $\subset$. Note that the lower proposition is a contradiction; it can never be true, but is a proper proposition.

3.2.2 Magnitudes with and without Order Relations

The general characteristics of measurements described in subsection 3.2 are intended to describe physical measurements in general, most of which are assumed to refer to magnitudes with continuously varying values. In this work, however, the focus is going to be put on finite sets of discrete outcomes, which, regardless of their importance in experimental physics (which is major, according to some authors like Friedkin [93]) covers most of the lexical measurements on which text IR is based.

Campbell's strict conception of measurement proves to be very limited when applied to social sciences, and is therefore not very popular in academic communities outside physics. In [94], for example, Joel Michell examines the history of such conception in psychology, with an emphasis on resolved and unresolved issues it has caused. Social sciences have pushed the development of radically new methodologies, motivated by the need to preserve qualitative measurement whilst overcoming its limitations. These methodologies differ substantially from those used in physics or "hard" sciences, because they are designed to make as few assumptions about the nature of empirical data as possible. In Grounded Theory [95], for example, the procedure begins with collection of data, and only similarity relations are needed amongst the collected observations (called *codes*) to build the structure of *concepts*, *categories*, and finally *theory*.

In this chapter totally qualitative measurements are not considered, since the proposed approach limits itself to the direct measurement of *lexical* attributes, and indirect determination of semantic attributes. The former can be stated in terms of natural numbers, and the second calls for more sophisticated mathematical concepts. In this work, we look for these concepts with more general relations in Quantum Theory, and, in particular, in the conception of measurement that has been developed to fit the logical requirements of this theory.

3.3 Measurement as Selection

In section 3.2.1, a description of the measurement process is finally described as definition of subsets of a set of all possible states: a subset for each outcome. They can overlap, and even some elements of the universal set can be in none of the subsets defined by measurements.

This description of measurement in terms of sets is already very powerful, and can in fact account for all measurement in classical mechanics [96]. However, there are still some problematic characteristics of semantic concepts that require further refinement for their adequate description. Logically problematic aspects of cognition [97] or the use of language by humans [18], amongst others, have been suggested to be more correctly described with concepts borrowed from Quantum Theory, which cannot be put in terms of sets and subsets, but in terms of Hilbert spaces and their subspaces.

To go beyond the set-theoretical description of measurement, it is useful to go one step back before, to a concept that Boole himself used, but abandoned in favour of a description in terms of classes (closely related to the one that is here referred to as set-theoretical) [98]

3.3.1 Boole's Selection Operators

A Selection Operator acts on a set of things, and selects only those fulfilling a condition. In the words of Boole himself ([99] , cited in [98])

> Let us conceive a class of symbols x, y, z possessed of the following character. The symbol x operating upon any subject comprehending individuals or classes, shall be supposed to select from that subject all the Xs which it contains[2].

An outcome of a measurement, then, can be represented as a proposition, and as such, it can define a selection operator, which in turn could produce a set by selecting elements. However, a selection operators can be defined in a slightly different way, so they can support certain non-Boolean features like measurements that interfere with one another. This takes us to a new formulation of measurement provided by Quantum Theory. This is not possible just working with sets, but using a more general version of what selection operators can generate, and a way for these more general operators to produce numeric outcomes from a measurement.

[2] Xs being defined as the elements having a common quality, or, equivalently, as we could say now, for whom a certain proposition holds as true.

3.3.2 The Concept of Measure

Formulating measurement in terms of propositions gives a formal logical framework ground to it, but for practical uses, we usually expect to get numbers from it. The concept that allows us to do so is *measure*

A *Measure*, in mathematics, is a systematic way of assigning numbers to a structure with a partial order relation (the subsets of a set is an example), such that the order relations are reflected in the numbers. This concept is the fundamental basis of the modern definitions of probability, and plays a paramount role in geometry and other branches of mathematics. Measure is also at the core of any definition of comparative and quantitative measurements, since it provides the formal bridge between the sets defined by selection operators and the numbers (or other ordered entities, in the case of comparative measurements) we use to represent the results.

Definition 3.1 (Measure)

A **measure** is a map μ from a lattice L_1 to a (usually simpler) lattice L_2, which preserves order relations.

$$\begin{gathered}\mu : L_1 \to L_2 \\ (a \geqslant b) \Rightarrow (\mu(a) \geqslant \mu(b))\end{gathered} \tag{3.10}$$

The second lattice L_2 is usually taken to be a completely ordered set, for example a set of real numbers.

Example 2

Measure: Let us consider all the subsets of a set $\{a, b, c, d\}$. Taking "is included in" $\subset$ and "includes" $\supset$ as order relations, we obtain a lattice. We can define a measure by assigning an integer number to each subset, by assigning positive numeric weights $\{w_a = 1, w_b = 2, w_c = 3, w_d = 4\}$ to the elements, and define a function μ that assigns a sum of weights of present elements to each subset:

$$\mu(S) = \sum_{\text{element } i \text{in } S} w_i \tag{3.11}$$

where S is a subset of $\{a, b, c, d\}$. For every pair of subsets S_1 and S_2 the order is preserved by the measure, meaning that

$$\forall S_1 \supset S_2, \mu(S_1) \geqslant \mu(S_2) \tag{3.12}$$

This example is shown in figure 3.3

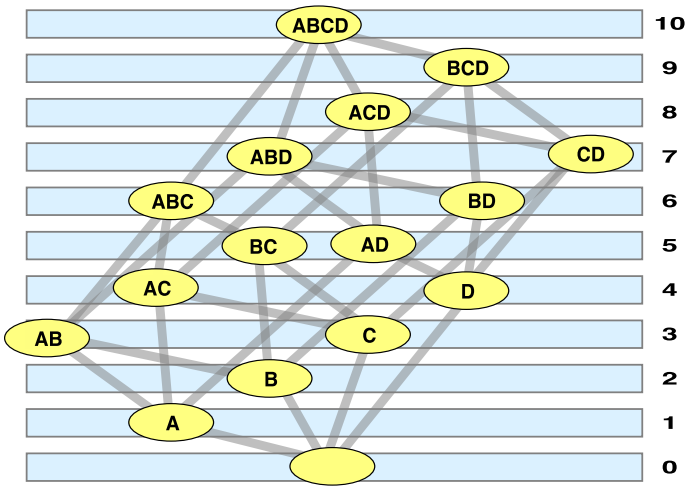

Figure 3.3: Example of a measure for a set of subsets of a set

The measure is defined as the sum of weights for the elements present: $w_A = 1, w_B = 2, w_C = 3, w_D = 4$. Each blue bar corresponds to a value of the mesure, and thick lines show inclusion relations. Relation (3.12) can be verified in this example, meaning that the sum of weights define a well-behaved measure

---

From the definition and example of measurement, it is clear that the assignation of a rational number (the magnitude of the measured observable) to the elements of the scale defined in law 2 of measurement, is in fact a measure.

There are two particular measures that have a paramount practical importance: **valuations** and **probabilities**.

Definition 3.2 (Valuation)

A **valuation** for a lattice is a map that assigns to each element, an element of another lattice of *valuations*[3] [92] (for example $\{true, false, true \geqslant false\}$) such that the order relations are preserved (a lattice homomorphism [101]). Some examples are shown in figure 3.4 for boolean lattices of propositions.

Definition 3.3 (Probability Measure)

A *Probability Measure* is a measure whose image set is the interval of real numbers between 0 and 1. The infimum is mapped into 0 and the supremum is mapped into 1.

[3] In other works the definition of valuation is narrower than ours; the maps assigns elements from a particular class of lattices: a chain, which has complete order (there is an order relation between any pair). Some authors even define valuation as a map to the set of real numbers [100].

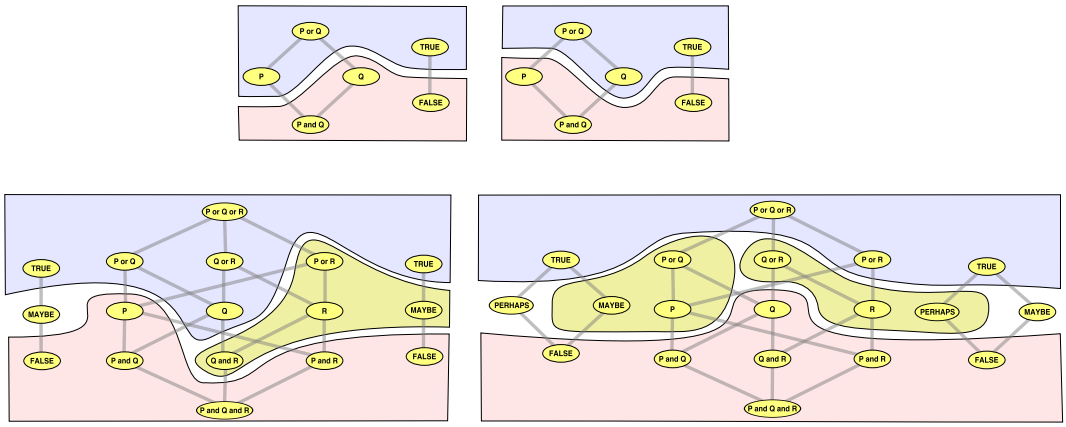

Figure 3.4: Some examples of valuation of lattices

The upper two are the usual $true/false$ valuation. The lower two are multi-valuations. These have also to comply with rules of the operations and and or: for example, in the left-hand graphic $maybe\ and\ true\ =\ maybe$, and $maybe\ or\ false\ =\ maybe$. Note that in the four-valued case, the lattice of valuations implies that $perhaps\ and\ maybe\ =\ false$ and $perhaps\ or\ maybe\ =\ true$, which sounds intuitively strange, but is formally coherent.

As we have shown in section 2.3 one of the ways of avoiding the rigidity of Boolean formulas for IR have been using probabilities and using fuzzy logics. Exotic non-distributive valuations have not been used yet in IR, but they might have a place in approaches where indeterminacy is allowed on truth values, as is the case in subjective logics [102, 103].

3.3.3 Boolean Algebras

In section 3.2.1 we used operations between propositions like *and* $\wedge$, *or* $\vee$ and *not* $\neg$ to describe outcomes of measurements. These elements, together with a lattice of propositions, form a **Boolean algebra** [92]. The elements and properties of this algebra correspond one-to-one to those of a similar algebra defined for sets, whose operations are the join, the meet and the complement with respect an universal set. Operations between sets can be defined in terms of elements x of the sets defined by the propositions as follows:

Definition 3.4 (Meet)

Meet of two sets:

$$[x \in (A \cap B)] \iff [(x \in A) \wedge (x \in B)] \tag{3.13}$$

Definition 3.5 (Join)

Join of two sets:

$$[x \in (A \cup B)] \iff [(x \in A) \vee (x \in B)] \tag{3.14}$$

Definition 3.6 (Complement)

Complement of a set with respect to a universal set U:

$$[x \in (U \backslash A)] \iff [(x \in U) \wedge \neg(x \in A)] \tag{3.15}$$

These operations are closely related to the order relations $\supset$ that define the lattice of sets, by the following relations:

$$(A \cup B) \supset B \qquad (A \cap B) \subset B \tag{3.16}$$

3.4 Non Compatible Measurements and Quantum Logics

A key characteristic of Quantum Theory is that it can deal with some observables whose observation inevitably introduces noise on the posterior results of the measurement of other observables. Let us suppose that A and B are binary observables, and they are incompatible. Measuring A would allow to assess the truthfulness of propositions asserting the two possible results a_1 and a_2. One is false, and the other is true. After having done so, however, the corresponding propositions about the observables of B cannot be said to be true or false. However, it still must be the case that $b_1 \vee b_2$ is true (B is still a binary observable) and $b_1 \wedge b_2$ is false (they are still mutually exclusive) [104]. This property lead Birkhoff and von Neumann in 1936 [105] to point at a particular logical law as the weakest link of Boolean logic as a basis for physics: **distributive law**.

3.4.1 Quantum Ideal Measurements and Distributive Law

Distributive law is a relation between two binary operations, that is probably most known as holding between multiplication and sum:

$$A \cdot (B + C) = A \cdot B + A \cdot C \tag{3.17}$$

where A, B and C are numbers, vectors or matrices, and $(\cdot)$ is any product defined between them.

The operations of Boolean algebras defined in subsection 3.3.3 fulfill a similar form of **distributive law** [92]:

$$A \vee (B \wedge C) = (A \wedge B) \vee (A \wedge C) \tag{3.18}$$

Since the failure to fulfill distributive law seems to be such a fundamental feature of incompatible observables, van Rijsbergen [10] uses the failure to comply to this law to actually *define* incompatibility:

Definition 3.7 (Compatibility)

measurement M_A is compatible to measurement M_B iff: The proposition that an outcome B was obtained in M_B can be decomposed in the case when any outcome A is obtained from M_A, and the case when an outcome excluding A ($\neg A$) is obtained.

$$B = B \wedge (A \vee \neg A) = (A \wedge B) \vee (\neg A \wedge B) \tag{3.19}$$

.

All observables that can be represented as operators partitioning sets will turn out to be compatible, since propositions about their results will fulfill distributive law [106]. When performing one measurement erases information from the other, they cannot be represented as simple partition of sets, and it becomes necessary to adopt a formal representation that allows for this non-distributive behaviour. However, it is also desirable to keep boolean logic as holding in a simple way in limited conditions (compatibility), and provide also a theoretically and operationally simple way of dealing with data. These goals are both accomplished by a view of measurement proposed within Quantum Theory, that is going to be discussed in next subsection.

3.4.2 Ideal Measurements according to Quantum Theory

To develop his axiomatisation of Quantum Theory, Lucien Hardy [107] proposed a view of measurement that is very appropriate for the purposes of this work. This scheme is based in three key concepts that are hereby defined:

Definition 3.8 (System)

System: is a part of the universe that is (conceptually or physically) isolated to be considered as the subject of any measurement. It is usually necessary to either have a fair amount of such systems to perform different procedures on them, or be able to use the same system again and again for a number of experiments.

Definition 3.9 (Ensemble)

Ensemble: is a large collection of systems that are composed of the same elements and prepared in the same way, but are otherwise independent to each other.

Definition 3.10 (State)

State: is a condition of a given system that is the result of a certain preparation, and produces results that are determined (up to a bounded statistical uncertainty). If two states are the same, the probability distribution for the results of every measurement on them should be the same.

Definition 3.11 (Measurement)

Measurement: Is the procedure where the value of a certain observable is assessed. This concept is going to be defined with more detail in this section.

Having defined the concepts involved, Hardy's scheme for measuring consists in the following steps:

1. An ensemble of a large number of similar copies of a physical system is obtained. The *systems* are assumed to be totally similar to each other, but the *states* of them can be allowed to be different to each other to a certain extent.

2. Some preparation procedure is performed on each system ($\{S_i\}$ in Figure 3.5) defining therefore a *state*.

   The state is defined only by the selections, transformations, etc. used in the preparation procedure. It can be, therefore, very vaguely defined, and is, for most theories, represented statistically.

3. Measurements ($\{M_i\}$ in Figure 3.5) are sets of detectors that act as **selection operators**, letting only a system that possesses a certain value for the measured characteristic pass, or rejecting it otherwise.

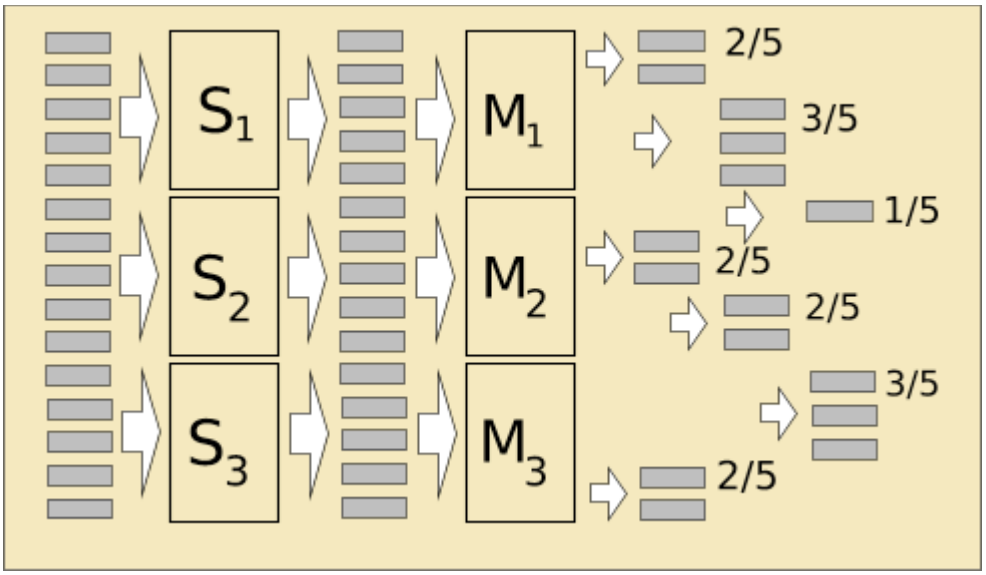

Figure 3.5: Hardy's scheme of measurement. In this example, measurements are sets of three detectors (selection operators). In the graphic, each state is allowed by exactly one filter, which is not always necessarily the case. At the right of the figure, the small arrows indicate which detector allowed which state, and the numbers indicate the fraction of the systems involved in the measurement that were preserved (chosen) by each one of the three filters

3.4.3 An Operator-Valued Measure

The mathematical object that represents this selection operators in Quantum Theory is the *projector* defined on a Hilbert space. A Hilbert Space is a vector space defined on the field of complex numbers, with an inner product.

Definition 3.12 (Projector)

A projector is a linear operator Π acting on a Hilbert space that fulfills the following conditions:

$$\forall|\psi\rangle,\langle\Psi|\Pi|\Psi\rangle \in \mathbb{R} \tag{3.20}$$

$$\forall|\psi\rangle,0 \leqslant \langle\Psi|\Pi|\Psi\rangle \leqslant \langle\Psi|\Psi\rangle \tag{3.21}$$

$$\Pi^2 = \Pi \tag{3.22}$$

Every projector defines a subspace. When a vector lies on a subspace, it is an eigenvector of the corresponding projector with eigenvalue 1:

$$(|\Psi\rangle \in S) \iff (\langle\Psi|\Pi_S|\Psi\rangle = \langle\Psi|\Psi\rangle) \tag{3.23}$$

When a vector is orthogonal to a subspace, it is an eigenvector of the corresponding projector with eigenvalue 0:

$$(|\Psi\rangle \perp S) \iff (\langle\Psi|\Pi_S|\Psi\rangle = 0) \tag{3.24}$$

The size (number of dimensions) of a subspace is to the *rank* of the corresponding projector:

Definition 3.13 (Rank)

The number of orthogonal vectors that are left invariant by a projector is its **rank**. The rank of a projector can be computed as its trace, which is always integer.

The *trace*, on the other hand, is defined as follows:

Definition 3.14 (Trace)

Given any basis set $|i\rangle$ that spans the whole space, the **trace** ($Tr(\cdot)$) is a linear functional that assigns a number to any linear operator. It is defined as follows:

$$Tr(A) = \sum_i \langle i|A|i\rangle \tag{3.25}$$

where $\{|i\rangle\}$ is a set of orthogonal vectors with norm one.

Sets can be represented with a set of commuting projectors and their corresponding subspaces. To represent a universal set U and its elements $\{e_i\}$, an equal number of orthonormal vectors $\{|i\rangle\}, \langle i|i\rangle = 1, \forall(i \neq j), \langle i|j\rangle = 0$, is chosen. The whole set would be represented by the whole space, and projectors would be operators that erase some of the elements if present. A projector representing subset S would be defined as follows:

$$\Pi_S = \sum_{e_i \in S} |i\rangle\langle i| \tag{3.26}$$

In the same way that there can a relation between elements and sets, there can be one between vectors and subspaces. This can be put in terms of projectors: if the vector remains unchanged under the action the projector (3.23), then it belongs to the subspace. When a vector is neither orthogonal nor contained in a subspace, we say that *part of it* lies on the subspace, or that there is an **overlap** between the vector and the subspace. Overlap is something that do not have an equivalent in sets: an element either belongs to a set, or to its complement; however, a vector can have a nonzero overlap both with a subspace and in its orthogonal complement.

3.4.4 Boolean-Like Algebras with Projectors

In the same way that in section 3.3.3 binary operations between sets can be put in terms of propositions about the membership of an element to a set $x \in A$, we can define them in a similar way with propositions about inclusion of a vector in a subspace:

Definition 3.15 (Meet of Projectors)

meet of two projectors

$$(\langle\psi|[A \cap B]|\psi\rangle = \langle\psi|\psi\rangle) \iff (\langle\psi|AB|\psi\rangle = \langle\psi|\psi\rangle) \tag{3.27}$$

where $|\psi\rangle$ is any complex vector in the space where projectors A and B operate.

Definition 3.16 (Join of Projectors)

Joinof two projectors

$$(\langle\psi|[A \cup B]|\psi\rangle = \langle\psi|\psi\rangle) \iff (\langle\psi|[A + B - AB]|\psi\rangle = \langle\psi|\psi\rangle) \tag{3.28}$$

where $|\psi\rangle$ is any complex vector in the space where projectors A and B operate.

Definition 3.17 (Complement of Projectors)

Complement of two projectors

$$(\langle\psi|[A\backslash B]|\psi\rangle = \langle\psi|\psi\rangle) \iff (\langle\psi|[A - B]|\psi\rangle = \langle\psi|\psi\rangle) \tag{3.29}$$

where $|\psi\rangle$ is any vector in the space where projectors A and B operate. A is usually taken to be as the whole subspace, and in that case, the complement with respect to A is simply called "complement".

For rank-one projectors that are not equal, the meet is always the null operator, since there is no vector that is contained in both one-dimensional subspaces, and the join is a rank two projector, that can be computed by the formula derived in appendix B:

$$|a\rangle\langle a| \cup |b\rangle\langle b| = \frac{(|a\rangle\langle a| - |b\rangle\langle b|)^2}{1 - \langle a|b\rangle\langle b|a\rangle} \tag{3.30}$$

The cases when the meet of projectors cannot be taken as the product, and the join cannot be taken as the sum minus the meet, can be recognised by the means of a logical relation between these two operations called *distributive law*

Definition 3.18 (Distributive Law)

Distributive law is fulfilled between two binary operations $\cup$ and $\cap$ when for any three objects A, B and C:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap B) \tag{3.31}$$

In the next section, we will discuss why distributive law is something that can (and perhaps should) be dropped in the logical framework of a lexical measurements scheme that aims to capture semantic contents.

3.5 A Logic of Projectors for Information Retrieval

Even though non-distributivity means that logics are not Boolean, it is easy to think in examples when it is violated when dealing with aboutness. Consider, for example, how a human would gather documents given some complex directions that combine assessment of the following topics:

$$\begin{aligned} A &= \text{D is about trees} \\ B &= \text{D is about computers} \\ C &= \text{D is about apple} \end{aligned} \tag{3.32}$$

The left-hand side of equation (3.31) would read:

$A \cup (B \cap C) =$"D is either about trees or computers and apple"

while the right-hand side, would be:

$(A \cup B) \cap (A \cup B) =$"D is both about tree or computers and about trees or apple"

Even though the expressions are equivalent in terms of sets, the way our subjects would look for the documents is likely to produce different results: for $A \cup (B \cap C)$ they would look for documents on apple computers, and then they would look for documents about trees, probably introducing a bias towards trees as a data structure coming from the previous query. For $(A \cup B) \cap (A \cup B)$, on the

other hand, they would independently look for a set of documents about trees and computers, and a different set of documents about apple trees; in this case the set query about apple trees would not be affected by any bias.

A suspicious reader will probably notice in this example that we are writing sentences like "(D is about A) and (D is about B)" as "D is about A and B", assuming that there is a map from the sets of documents D and the sets of topics A and B, that allows to translate an "and" operation between propositions about documents to an "and" relation between topics. That is, in fact, a basic assumption of this work, namely:

Assumption 2

Aboutness is a relation that maps a set of documents to a topic. Topics are not representable as sets themselves, because meaningful logical operations between them would not be distributive, but can be represented as subspaces. Assessments of aboutness do not behave as partitions of a set of documents as Boolean selection operators, but as quantum ideal measurements, and lexical measurements on them should reflect this fact.

3.6 Conclusions: How the introduced concepts will be used

In this chapter, we have presented a formal account of measurement, and what its logical basis is. This basis, is a propositional logic (understood as a lattice of relations, plus a set of binary operations and an interpretation scheme) is presented in its Boolean (classic) form, showing how this can be different whenever incompatible measurements are considered. This paves the way to build a quantum-like logic of lexical measurements, that can be used to describe and retrieve natural language text documents. In the next chapter such lexical measurement is proposed, its properties are discussed, and some practical applications of the concepts are outlined.

Chapter 4

The Selective Eraser (SE)

In chapter 3, we showed how the notion of measurement is conceived and described in QT. We proposed a way in which sets of measurements and propositions about their outcomes can be given a logical structure, and how this structure can be slightly different to that accounted for by Boolean logics, but coincides with that used in QT. In this chapter, we introduce a transformation that mimics ideal quantum measurements in its mathematical properties: the Selective Eraser (SE).

We start by defining SEs in section 4.1. In section 4.2 the behaviour of SEs is examined from the point of view of Campbell's three laws of measurement (introduced in section 3.2). As these laws call for a mapping from measurements to numbers, in section 4.2.2 we define different kinds of norms for documents, which can be used to define similarity measures for them. In 4.3 we outline several ways of examining text documents with SEs, and interesting links are found with existing techniques to process text for IR. In section 4.4, order relations between SEs are further explored, together with the elements they provide for the description of text. In section 4.5 we discuss how exhaustively a document can be represented by the lexical measurements introduced, and how this representation can be more exhaustive than that used in bag-of-words approaches. In section 4.6 we show how the proposed measurement scheme deviates from Boolean logic, and how the obtained non-Boolean relations are similar to those found in Quantum Logics. In section 4.7, the role of probabilities on the proposed scheme is discussed. In section 4.8 we develop the parallel between text transformations (SEs) and abstract linear operators, to define a linear algebra that allows us to build quantum-like operators acting on text that behave like linear operators acting on

a Hilbert space. Representations of both measurements (SEs) and states (documents and queries) are discussed in section 4.9, where a link is established to the logical uncertainty principle, wich could lead to a new formulation of non-boolean, logic-based IR systems. Finally, in section 4.10 the contents of the chapter are summed up.

4.1 Definition

Chapter 3 began with a remark on how, when browsing a newspaper, a user's attention would be captured by a handful of keywords, and ten the user is likely to select the surrounding text to examine it more closely. Then, in section 3.3 whe showed how measurement can be seen as a selection procedure. These two ideas lead quite directly to a definition of a lexical measurement that we will call the Selective Eraser (SE).

SEs, first of all, are operators that select part of a text document. The definition of a SE that we will use is one of the published results of this work [108]:

> **Definition 4.1 (Selective Eraser)**
> A **Selective Eraser** (or simply *Eraser*) is a transformation $E(t, w)$ which erases every token that does not fall within any window of w positions around an occurrence of term t in a text document. These Erasers act as transformations on documents producing a modified document with some erased tokens, much as projectors act on vectors or other operators. Erased tokens will have an undefined identity: they could be any term but t (the central term).

This concept was first introduced in [109]. Here some of its properties are shown, specially those that resemble ideal measurements as described in Quantum Theory. The concept of SE is here presented in a more elaborated way, introducing tokens whose identity is not fully determined. There are two key aspects of SE that allow to build a powerful lexical scheme upon them:

- They select part of the text in the document which is close to occurrences of a term. This preserved text is likely to be part of information units in which the concept represented by

the central term is involved in some way; some SEs could capture complete information units (like sentences)

- When they are applied for the first time, information outside their scope is lost, but no further information is lost when applied for a second time. This is called idempotency, and can be expressed as:

$$\forall D, E(t,w)[E(t,w)\cdot D] = E(t,w)\cdot D \quad \text{or simply} \quad E(t,w)^2 = E(t,w) \tag{4.1}$$

Figure 4.1: Action of different Selective Erasers on a text sequence (document) D

The lower sequence represents the transformed sequence. The letters with a tilde represent an undetermined letter that can be anything but the central letter

To understand the action of a SE on a sequence of terms, it is useful to consider for each position, the probability distribution that a particular term occupies the position. We can consider the initial text as having a defined term occupying every position, so all of the distributions would have only one non-zero entry; a probability of 1 for a term. After the application of the SE, the unerased text would remain completely determined, but in other positions the term that is there becomes uncertain; now the only thing that is known is that the term is not the central term (this is represented in figure 4.1 by a term with a tilde on top, meaning "not this term". The distributions of probabilities for the positions occupied by "not this term" are almost flat: i.e. they are flat except for a one term (the central term of the SE) with probability of 0. The non-zero probabilities in this almost flat distributions are $\frac{1}{N_V-1}$ where N_V is the size of the vocabulary.

4.2 SEs and the Laws of Measurement

In section 3.2 we discussed the laws of measurement. These laws can be seen as requirements a formal definition of measurement should fulfil. Campbell's three laws deal, respectively, with Ordering (definition 3.1), Additivity (definition 3.2) and Changes of Scale (definition 3.3).

To address the first law (ordering), we will relate SEs to the set relations involved in measurements, including Boolean join (union) meet (intersection) and set inclusion, which will allow the definition of order relations between SEs. The second law (additivity) has to do with counting, and we will approach this issue through the quantification of information in the sequence of terms of a document. Finally, the issue of scales (third law) will be considered by discussing the fine-grained or coarse-grained character of lexical measurements.

4.2.1 First Law: Boolean Operations and SEs

In chapter 3 the results of measurements were represented as propositions, in order to build a mathematical representation of measurements. Our first step to describe the action of SEs in documents as the results of measurements, will be to consider propositions referring to every token on the transformed document as: $O_t(D_i) =$ "token i in D is t" or $\neg O_t(D_i) =$ "token i in D is not t". We can then define Boolean meet $\cap_B$ and join $\cup_B$ of two SEs, defining them as:

Definition 4.2 (Boolean Meet (Intersection))

The **Boolean Meet (intersection)** of two SEs is the transformation that preserves the information preserved by both SEs:

$$\forall t, O_t(([E(a, w_a) \cap_B E(b, w_b)]D)_i) = O_t((E(a, w_a) \cdot D)_i) \wedge O_t((E(b, w_b) \cdot D)_i) \quad (4.2)$$

This means that two propositions are equivalent: one is "*a term is in a particular position after applying the Boolean meet*" and te other is "*the term is in that particular position after applying one SE,* ***and*** *it is in that position after applying the other*" Positions of the document erased by this operation would be undetermined, except for the fact that they cannot be either of the central terms.

Definition 4.3 (Eraser Boolean Join (Union))

The **Boolean Join (union)** of two SEs is the transformation that preserves the information preserved by any of the SEs:

$$\forall t, O_t(([E(a, w_a) \cup_B E(b, w_b)]D)_i) = O_t((E(a, w_a) \cdot D)_i) \vee O_t((E(b, w_b) \cdot D)_i) \quad (4.3)$$

This means that two propositions are equivalent: "*a term is in a particular position after applying the Boolean join*" and "*the term is in that particular position after applying one SE,* ***or*** *it is in that position after applying the other*"

As it is also the case with the Boolean meet, positions erased by the Boolean join are undetermined except for the fact that they cannot be either of the central terms.

Another Boolean operation that can be defined is the complement of a SE.

Definition 4.4 (Eraser Complement)

complement of a SE $\neg E(t, w)$: Is the transformation that erases every token that is preserved by the SE, and preserves every token that is erased by the SE.

Terms erased by this operation are left totally undetermined; they could be any term at all, including the central term of the eraser t.

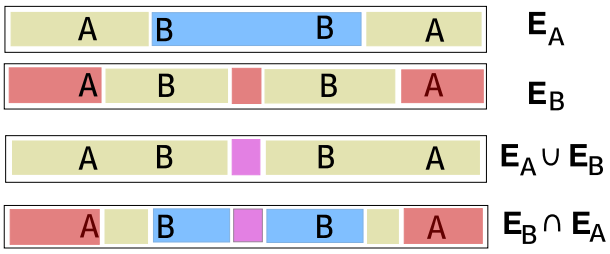

Figure 4.2: Boolean Meet and Join of two SEs

Text with light background remains unchanged; tokens with lightly-dark (or blue if in colour) background are indetermined but different to A, tokens with heavy-dark (or red if in colour) background are undetermined but different to B. Tokens with medium-dark (or magenta if in colour) background are undetermined but different to both A and B, and tokens with dark magenta background are totally undetermined.

In figure 4.2 we can see, in an example, how there can be two kinds of undetermined terms in the case of the meet: those that are not one of the central terms, and those that are not any of the two

central terms. In terms of the probability distribution of terms, the maximum probability in the distributions for some positions will be 1, in others $\frac{1}{N_V-1}$ and in others $\frac{1}{N_V-2}$.

Intuitively, we can imagine how a SE will erase more than other, when the windows of text preserved by one are contained in the windows preserved by the other. When the text preserved by $E(t_1, w_1)$ is also preserved by another $E(t_2, w_2)$, then we can say that the second includes the first $E(t_2, w_2) \geqslant E(t_1, w_1)$.

If we can compare unerased text in documents, we can define inclusion in terms of the join (union) of SEs:

Definition 4.5 (Inclusion Relation Between Erasers)

Inclusion relation between two SEs: A SE includes another, when the text preserved by the first is the same that is preserved by their join, when applied to the documents in a set C:

$$(E(t_1, w_1) \geqslant_C E(t_2, w_2)) \iff (\forall D \in C, [E(t_1, w_1) \cup_B E(t_2, w_2)]\ D =_u E(t_1, w_1)\ D) \tag{4.4}$$

where partial equality holds between two transformed documents $D_1 =_u D_2$ when their unerased tokens are equal. When the relation holds for any possible document, the relation is not referred to a set of documents C (as in $\geqslant_C$) but is simply an absolute relation $\geqslant$.

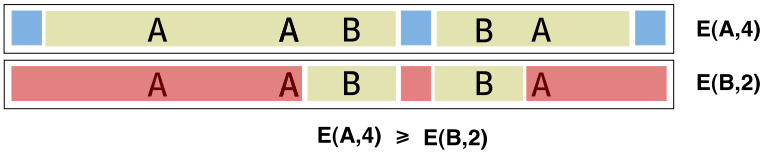

Figure 4.3: A SE including another.

$E(A, 4)$ preserves everything that is preserved by $E(B, 2)$. The join preserves the same text as $E(A, 4)$, but the unpreserved text is slightly different: for $E(A, 4)$ is constituted by tokens that are "not A" (magenta+blue) while for $E(A, 4) \cup_B E(B, 2)$ it is constituted by tokens that are "neither A nor B" (magenta).

This order relation can play the role of the inequalities involved in length measurements in section 3.2. A first parallel between SEs and measurements, could be as follows:

Consider the following result of a length measurement: $4cm \geqslant L \geqslant 3cm$. The result is defined by an order relation between two standards ($4cm$ and $3cm$) and a measured length L. In the same

way, we can situate a lexical measurement $E(t, w)$ in a measurements order between two standards $E(a, w_a)$ and $E(b, w_b)$ for a set of documents D. such that $E(a, w_a) \geqslant_D E(t, w) \geqslant_D E(b, w_b)$ This is not immediately useful, but in section 4.4 some applications will be discussed.

4.2.2 Second Law: Information and the Number of Unerased Terms

The parallel drawn between the measurement of length with a ruler and a measurement on a document with SEs is already a valid one, but it has some issues. The Second law of Measurement (definition 3.2) states that measurement should involve additive quantities. Additivity is used to define scales: adding units of a standard, an integer-numbered scale is made. In the last section, order relations are established between SEs, not between documents. If we want to define the usual lexical measurements, like counting occurrences on a document, we need to define a map that takes from documents to numbers. Maps like this are called **norms**.

Since what an eraser does on a document is precisely erasing information, it is quite a natural choice to define Shannon information (defined in section 2.4.2) as a measure for documents. For a given position in the text, the least informative situation would be not having a clue about which term occupies it; This will be not having any information about the identity of that term. This would correspond to a flat probability distribution through all the vocabulary; one where every probability is $\frac{1}{N_V}$ (where N_V is the size of the vocabulary). The largest amount of information we can have about this position in the text is knowing exactly which term occupies it; this would correspond to a distribution where one term has probability 1 and the rest have probability 0. If we adopt a scale for information where every determined position has 1 unit, and an undetermined position has 0 units, we define information as follows:

$$I_{\text{position } x} = \frac{\log(N_V) + \sum_t P(t \text{ is in position } x) \log(P(t \text{ is in position } x))}{\log(N_V)} \tag{4.5}$$

Considering all the positions of a document as independent, the total information of the document in these information units would be precisely the length when all of its terms are determined. However, when a SE has acted on it, some of the positions would have an undetermined term, so that the only thing we know about them is that the term occupying them is not the central term of

the SE. This means that the amount of information is:

$$I_{\text{erased position}} = \frac{\log(N_V) + (N_V - 1)\frac{1}{N_V - 1}\log(\frac{1}{N_V - 1})}{\log(N_V)} = 1 - \frac{\log(N_V - 1)}{\log(N_V)} \tag{4.6}$$

The total information in a document after applying a SE would then be simply:

$$\begin{aligned} I(E(t,w)D) = N_u + (L_D - N_u)\left(1 - \frac{\log(N_V - 1)}{\log(N_V)}\right) = \\ = N_u\left(\frac{\log(N_V - 1)}{\log(N_V)}\right) + (L_D - N_u)\left(1 - \frac{\log(N_V - 1)}{\log(N_V)}\right) \end{aligned} \tag{4.7}$$

where L_D is the length of the document, N_V is the size of the vocabulary and N_u is the number of unerased tokens.

Fraction $\frac{\log(N_V - 1)}{\log(N_V)}$ is almost 1 for medium to large vocabularies, so information is almost a preserved token count. For a vocabulary of only 10 terms, it would be 0.0458, for 100 terms it would be 2.10×10^{-3} and for a typical vocabulary of 10^5 terms, it would be 8.69×10^{-7}. As a good approximation, we can say that information counting in this scale is almost equivalent to a simple count of the unerased tokens:

Definition 4.6 (Unerased Token Counting)

Unerased Token Counting is the number of tokens that have not been erased by a SE, i. e. whose identity is still determined. It is represented by vertical bars on the sides $|\cdot|$.

The length L of a document D that has not been transformed by any SE would be precisely $L_D = |D|$. An occurrence count of a term t can be obtained with this norm as well:

$$N_{t \text{ in } D} = |E(t,0) \cdot D| \tag{4.8}$$

It is clear that any token could be used as a unit, and a scale can be produced adding repetitions of it, for example D_1 =“unit”, D_2 =“unit unit”, D_3 =“unit unit unit” and so on. Comparison of the norms with a scale $\{D_i\}$ would be the formal way of defining token counting. It is also possible to define other kinds of norms. If, for example, the occurrence of different terms brings a different amount of information (an intuitive assumption) extra considerations can be included that are additional (and perhaps independent) than that of whether the term in a particular position is determined or not. Norms coming from this consideration are explored in section 4.2.2

As it was mentioned before, a norm $|\cdot|$ is a systematic assignation of a real number to an object (e.g, for documents $|D| \in \mathbb{R}$) and can provide the link between a measurement procedure and a numeric result.

A natural information-based measure was defined for documents, which considered all terms equally important and independent, and how it would be desirable to relax these assumptions. With that in mind, is possible to make use of SE to define a family of norms, called *Term-Weighted Norms*:

Definition 4.7 (Weighted Norm)

Given a set of term weights $\{\alpha_t\}$, the corresponding **Term-Weigthed Norm** $|\cdot|_{\{\alpha_t\}}$ is defined as:

$$|D|_{\{\alpha_t\}} = \sum_t \alpha_t |E(t,0) \cdot D| \tag{4.9}$$

This norm can be defined with any term weighting scheme. A key principle in indexing is some key terms that convey more information about the topic of a document than others, and this has been formalised in concepts such as *term specificity* [42]. Some approximations to the quantification of term specificity can also be expressed in terms of SEs, like IDF:

$$IDF(t) = \log\left(\frac{|D|}{\sum_{D\in C} \frac{|E(t,\infty)\cdot D|}{|D|}}\right) \tag{4.10}$$

A problem that arises when using this norm, is how to evaluate the norm of a term that has been transformed by a SE (for example $|E(t,w)D|_{\{\alpha_i\}}$) is that the evaluation implies that two SEs are applied one after another:

$$|E(t,w)D|_{\{\alpha_i\}} = \sum_{t_i} \alpha_i |E(t,0)[E(i,w)D]| \tag{4.11}$$

To this point, this composition of SEs (applying one after another) has not been defined, but it suffices to say that there is nothing to it that can make the definition of these norms problematic.

Definition 4.8 (Product of Erasers)

The Product of two SEs is the transformation that results of applying the first and then the second. The application of the first SE will mark some tokens of the document as erased, and for the second eraser only unerased occurrences of its central term are clear-cut; erased tokens could be its central term of the second eraser with a probability of $\frac{1}{N_V}$. The second SE will

then mark further tokens as erased. This product is, in general non-commutative as is shown in figure 4.4.

$$[E(t_2, w_2) \circ E(t_1, w_1)] \cdot D = E(t_2, w_2) \cdot n[E(t_1, w_1) \cdot D] \tag{4.12}$$

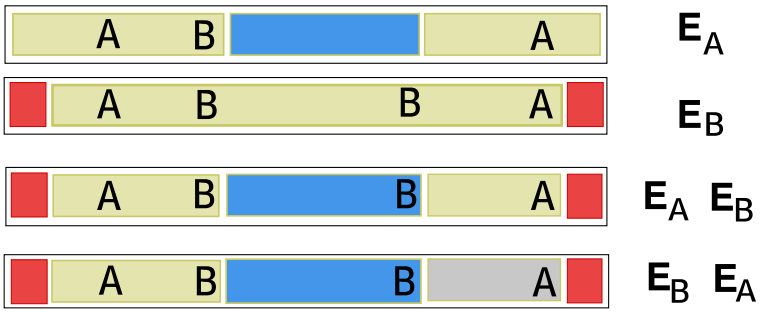

Figure 4.4: Products of two SEs in different orders.

Applying E_B first, as in the first product $E_A E_B$, all occurrences of A are preserved, but if E_A is applied first, as in $E_B E_A$, one of the occurrences of B is erased, resulting in the end in the erasure of the nearby occurrence of A as well. Areas in blue represent in determined tokens that are not A, and areas in red represent undetermined tokens that are not B. The difference between the product in two different orders is the area coloured in gray where the tokens could be either A or B, because there is a probability of $\frac{1}{N_V-1}$ that they were not erased.

In the figure it is also shown that when a SE finds a token that *could be* its central term with a certain probability, it would also erase with that probability, producing a zone where terms are semi-determined. If two SE are applied, for example $[E(a, w_a) \circ E(b, w_b)]D$ then SE $E(b, w_b)$ will erase first, and will leave positions where it is only known that term b does not occur, but a could occur, with a probability of $\frac{1}{N_V-1}$. SE $E(a, t_w)$ can find unerased text that is in the vicinity of these undetermined terms, and would then erase with a probability of $\left(1 - \frac{1}{N_V-1}\right)$. In the positions where this happens, the probabilities of different terms will have the following values:

- $\frac{N_V+1}{(N_V-1)^2}$ for the term that was initially in that position
- $\left(\frac{N_V}{N_V-1}\right)\left(\frac{N_V-3}{(N_V-2)(N_V-1)}\right)$ for all terms different to the initial term but also to b
- 0 for b.

The information in these positions will be an intermediate one between the almost flat probability with two values, and the certain probability. However, the information it contributes for large

vocabularies is very similar to that of the totally undetermined: almost none. Some values of the information are shown in table 4.1 Since preserved terms contribute with 1, and terms erased by

N_V	$I_{\text{undetermined}}$	$I_{\text{semi-undetermined}}$
10	4.5757×10^{-2}	4.7020×10^{-2}
10^2	4.5757×10^{-2}	4.7020×10^{-2}
10^3	2.1824×10^{-3}	2.1829×10^{-3}
10^4	1.4484×10^{-4}	1.4484×10^{-4}
10^5	8.6859×10^{-7}	8.6859×10^{-7}

Table 4.1: Contribution of a semi-determined position in the text to overall document information N_V is the size of the vocabulary, $I_{\text{undetermined}}$ is the information contained in a totally undetermined token, and $I_{\text{semi-undetermined}}$ is the information contained in a position that would be preserved with a probability of $\frac{1}{N_V-1}$ and erased with a probability of $\frac{N_V}{N_V-1}$.

one or both SEs in a product would contribute with nearly zero, definition 4.7 of weighted norm will be almost equivalent to just counting occurrences of a document and summing them with term-dependent weights.

4.2.3 Third Law: Scales and Units of Measure

Since occurrence count involve a quite natural unit of measure (the preserved token), natural numbers are enough to describe such count. The definition given of a norm for documents in the last section was devised to comply with this intuitive characteristic, taking the unerased token as the unit of occurrence. However, for more sophisticated lexical measurements on text documents, there is a wealth of analysis power that can be gained by considering diverse levels of coarse or fine-graining. One potentially useful aspect of the scheme of measurement proposed by this work has to do with the width factor w in SEs: it determines the size of the segments of the document sequence that are going to be preserved around occurrences of the central term. Considering families of SEs with different width factors introduces a rich structure of inclusion relations between SEs, and useful measurements of the sequence structure can arise from these relations, as will be discussed in section 4.3.1 for families of SEs with the same central term, and in section 4.4 the link of this with the analysis of co-occurrences of different terms will be established.

Ordering relations between SEs centred on different terms form a structure called *lattice* (to be defined and explained in section 4.4.2. These structures resemble those formed by different scales

(discussed previously in section 3.2.1) but differ in a subtle but key aspect. This key aspect, non-distributivity, will suggest the use of a mathematical representation brought from QT, called Positive Operator Valued Measure (POVM). This will be discussed in section 4.4.

4.3 Examining text with Erasers

With the definitions given up to this point, it is already possible to use SEs to obtain useful information from text. It is possible to put lexical measurements on the number of occurrences of terms in collections, as well as on relative positions of some occurrences with respect to others, in terms of different schemes based on SEs.

4.3.1 Term Frequency and Burstiness

Term frequencies can be obtained in a very simple way by applying SEs to a document:

$$TF(t \text{ in } D) = |E(t,0)D| \tag{4.13}$$

Just as in the case of the document norm, there is a family of modified term frequencies that can be defined with non-null width factors. These would be approximately proportional to term frequency for small w, but will increase at a sub-linear pace for higher values of w, because terms in the overlap between n windows will not be counted n times, but only one. This quantity can be bounded from above and below by considering two extreme cases: that with all the occurrences together, and that with all the ocurrences maximally and equally spaced in the document:

$$2w + |E(t,0)D| \leqslant |E(t,w)D| \leqslant (2w+1)|E(t,0)D| \tag{4.14}$$

Some probabilistic document rankings are logarithms of a ratio of probabilities: that of the occurrences being a sampling of a distribution that is characteristic of **elite** documents (documents about the topic), and that of occurrences being a sampling of a distribution characteristic of **non-elite** documents (documents not about the topic). If all subsequent occurrences of documents are independent, then such ranking would have the form:

$$R(D, topic) = \sum_{t \text{ in } D} \log \left(\frac{\prod_{\text{occurrences}} P(t \text{ in elite})}{\prod_{\text{occurrences}} P(t \text{ in non-elite})} \right)$$

$$\overset{\text{independence}}{=} \sum_{t \text{ in } D} \underbrace{TF(t \text{ in } D)}_{\text{document}} \log \left(\underbrace{\frac{P(t \text{ in elite})}{P(t \text{ in non-elite})}}_{\text{topic and collection}} \right) \tag{4.15}$$

In this way, for each term in the document there is a logarithmic factor that depends on the collection and topic distributions, and term frequency appears as a linear factor. This makes computation much easier, but the independence assumption is well known not to be valid in natural language, where terms are known to appear in bursts (the tendency to do so is called *burstiness*) [110].

An even simpler approach is to assume that the simple presence of a term is as good evidence of a document being about a subject, as is done in the so-called binary models; in these, term frequency is simply replaced by a 1 if the term is present or a 0 if it is not. This model, in spite of its naïve simplicity, has also a reasonable performance [19]. However, common sense seems to suggest that something in the middle ground between these two extreme models could probably work better.

It was noted by Wu and Roellecke [111] that considering only presence or absence amounts to *subsume* all subsequent occurrences of a term into a less restrictive event like "the term occurred". It is also possible to regard subsequent occurrences of a term as *semi-subsumed* events. In this way, Wu and Roellecke obtain an expression for documents with a fixed length that is formally very similar to that of the successful model BM25, proposed by Robertson and Walker as a rather heuristic twist of the 2-Poisson model in [112].

$$\begin{aligned} R_{semi-subsumed}(D, topic) &= \textstyle\sum_{t \text{ in } D} \underbrace{\frac{2\, TF(t \text{ in } D)}{TF(t \text{ in } D) + 1}}_{\text{document}} \log \left(\underbrace{\frac{P(t \text{ in elite})}{P(t \text{ in non-elite})}}_{\text{topic and collection}} \right) \\ R_{BM25}(D, topic) &= \textstyle\sum_{t \text{ in } D} \underbrace{\frac{(K_1 + 1) TF(t \text{ in } D)}{TF(t \text{ in } D) + K_1 \frac{L}{L_{avg}}}}_{\text{document}} \log \left(\underbrace{\frac{P(t \text{ in elite})}{P(t \text{ in non-elite})}}_{\text{topic and collection}} \right) \end{aligned} \tag{4.16}$$

where L is the length of the document and L_{avg} is the average length of the documents in the collection (they are the same if the length is fixed).

In quantitative terms, it could be said that the overlap between windows limiting the value of $|E(t,w)\cdot D|$ in equation (4.14) is a form of semi-subsumption, since part of the token counting can be attributed to several occurrences of the central term t when w is large enough.

Just as the semi-subsumption model would reproduce a BM25 term frequency factor with $K_1 = 1$, a wide-window token count would reproduce it with a different constant. To check this possible similarity between the functional dependence of BM25 ranking and that of wide-window measurements on term frequency, it is possible to translate the former to a linear dependence with a transformation of variables. Taking the inverses of term frequencies and the inverse of wide-window measurement will do the trick: let $x = |E(t,0)\cdot D|$ and $y = \frac{|E(w,t)\cdot D|}{2w+1}$. The linear relation between the inverses is as follows:

$$\frac{1}{y} \approx a + b\frac{1}{x} \quad \Rightarrow \quad y \approx \left(\frac{1}{a+b}\right)\left(\frac{\left(\frac{b}{a}+1\right)x}{x+\frac{b}{a}}\right) \tag{4.17}$$

Comparing with BM25 term-frequency-in-document function, the parallel becomes clear:

$$F_{BM25}(t \text{ in } D) = c\left(\frac{(k_1+1)\,x}{x+k_1}\right) \tag{4.18}$$

where x is the term frequency, c is an irrelevant proportionality factor, and k_1 is an adjustable constant.

An analogy to BM25 constant k_1 can be then obtained from a simple linear regression between the inverse of normalised wide-width count $\frac{|E(w,t)\cdot D|}{2w+1}$, and term frequency $|E(t,0)\cdot D|$.

In graphic 4.5 values for K_1 can be seen for chapter 2 of "The Origin of Species by Means of Natural Selection" [113], a text with 12323 tokens that are occurrences of 2104 different terms, for width factors from 0 to 1000. The relation between the computed k_1 and w is nearly a power law with negative exponent, with some deviation of this form for low values of w.

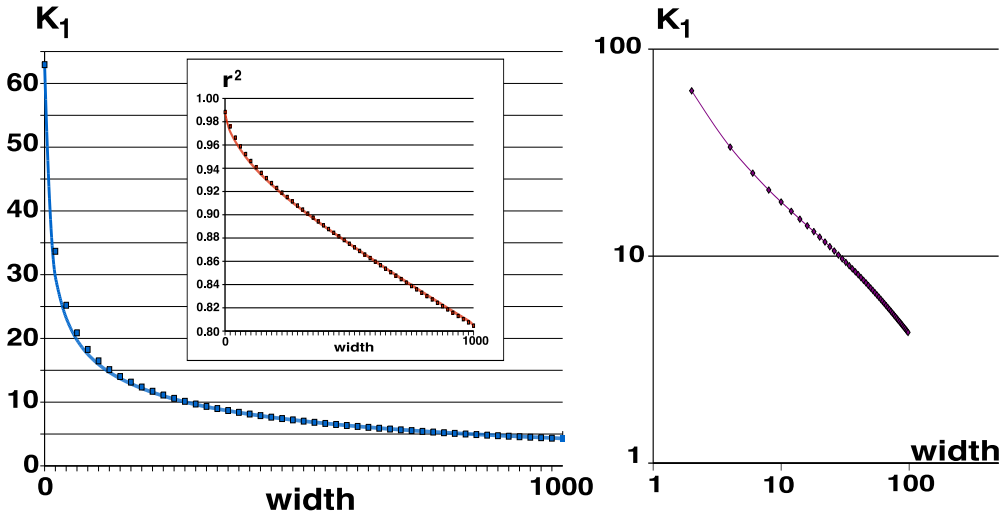

Figure 4.5: Values of constant k_1 for a BM25-like term function formula defined with SEs, as a function of width parameter w (on the right, in logarithmic co-ordinates).

This was obtained by linear regression between the values of $\frac{1}{|E(t,0)\cdot D|}$ and $\frac{2w+1}{|E(t,w)\cdot D|}$. Squared correlation coefficients are shown as well. Data was obtained from chapter 2 of Darwin's "*The Origin of Species by Means of Natural Selection*"

4.3.2 Distribution of Distances between Occurrences of a Term

Comparing the preserved terms counting for different widths is a way of finding out the distances between occurrences of the terms. To explain how this can be done, let us first consider a text so long that its borders will not influence the counting at all. In this document, each occurrence of a term T is much further from the borders than from any other occurrence. In such a document, $|E(t, w+1)\ D$ will be counting of tokens within the windows around every occurrence of T, minus the overlaps (since each token can only be counted once):

$$|E(T, w)\ D| = (2w + 1)|E(T, 0) \cdot D| - Overlap(T, w) \tag{4.19}$$

For two windows centred in different occurrences of T to overlap, it is necessary that the distance between the two occurrences is less than w, as is shown in figure 4.6

As we increase w from zero, the overlap increases in the number of distances between occurrence

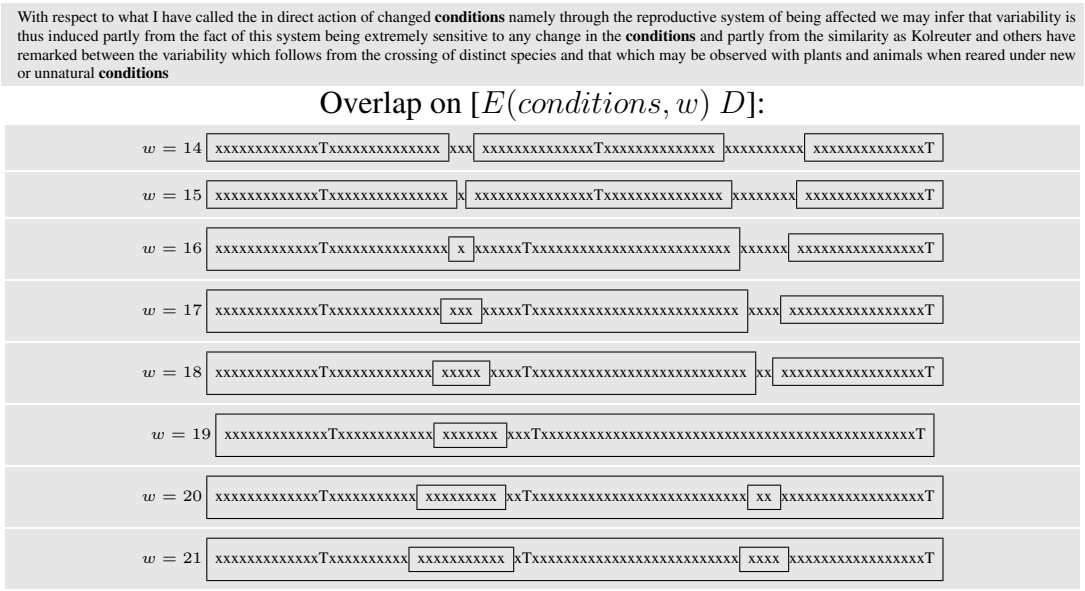

Figure 4.6: Overlap between windows centred on occurrence of term T

The distances between occurrences are 31 and 38 tokens. When $2w \leqslant 31$ there is no overlap; when $31 < 2w \leqslant 38$ there is one growing overlapping region, and when $38 < 2w$ there are two growing regions of overlap

that are less than w

$$Overlap(T, w) - Overlap(T, w-1) = \sum_{d_i} \sigma(2w - d_i) \tag{4.20}$$

where $\sigma(x)$ is the step function, with a value of 0 for $x < 0$ and a value of 1 for $x \geqslant 0$. This allows us to obtain an expression for the overlap at any value of w:

$$Overlap(T, w) = \sum_{d_i < w} (2w - d_i) \tag{4.21}$$

The exact token counting after applying $E(T, w)$ would be:

$$|E(T, w)\ D| = (2w + 1)|E(T, 0) \cdot D| - \sum_{d_i < w} (2w - d_i) \tag{4.22}$$

To include border effects, the distances between each border and the closest occurrence of the term has to be included, but **multiplied by 2**. This is because while erased sections between two occurrences of the term decrease by two (one window at each side), erased sections in the border

decrease just by one (one growing window on one side, the fixed border of the document in the other.

The rate of change of $|E(T,w)\cdot D|$ with w is changed by 2 when one of the distances between occurrences d_i is surpassed by $2w$. For $(2w+1) < length$, the number of distances lower than $2w$ can then be computed as:

$$N(d<2w) = \left((2w+1)|E(T,0)\cdot D| - |E(T,w)\cdot D|\right)/2 \tag{4.23}$$

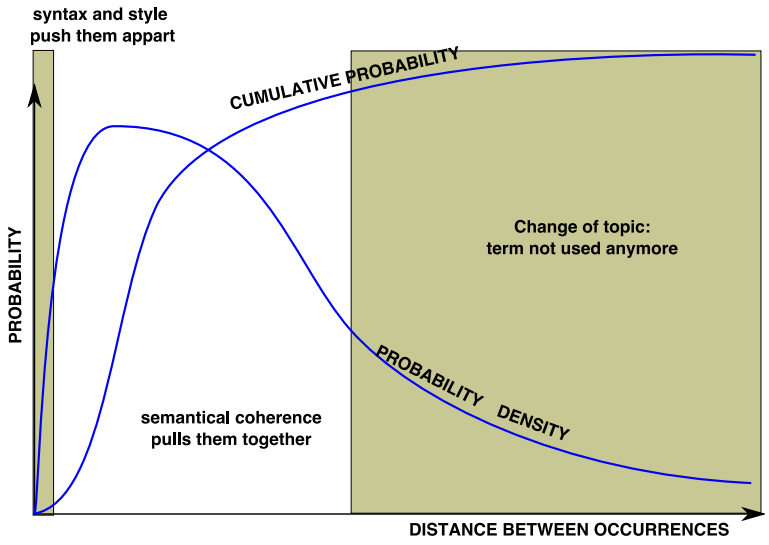

Figure 4.7: Distribution of distances between occurrences of a topic-related term

It is a known fact in natural language, that subsequent occurrences of terms at different distances show two opposite tendencies in different scales: in short scales, syntax and style tend to separate occurrences, but in long scales, topical coherence tends to keep them together [114]. A distribution of distances should then present the form shown in figure 4.7. Such form of the cumulative distribution can be compatible with a relation similar to (4.17), namely:

$$\left(\frac{1}{|E(T,w)\cdot D|}\right)^{\alpha_T} \approx \left(\frac{2w}{(2w+1)length}\right)^{\alpha_T} + \left(\frac{1}{(2w+1)|E(T,0)\cdot D|}\right)^{\alpha_T} \tag{4.24}$$

where α_T is a parameter that will presumably be different for each term. It has the effect of making the curve approach the flat asymptote faster (for large values of α).

Substituting (4.24) in (4.23), an expected form for the cumulative distribution of distances between

occurrences of term T is found:

$$N(d_T < (x-1)) \approx N_T x \left(1 - \frac{length}{\sqrt[\alpha_T]{((x-1)N_T)^{\alpha_T} + length^{\alpha_T}}}\right) \tag{4.25}$$

where $N_T = |E(T,0) \cdot D|$ is the number of occurrences, α_T is a parameter to be found for each term, and $x = 2w + 1$ is a dummy variable arising from the width of the SE used.

4.3.3 Choosing Keywords

According to several studies, like [115] and [116], distances between terms can be the used as a criterion for finding keywords. Using a modified version of (4.27) it is possible to select terms with inter-occurrence distance in a particular interval:

$$N(2w_{min} < d < 2w_{max}) = (2(w_{max} - w_{min})|E(T,0) \cdot D| - |E(T,w_{max}) \cdot D| + |E(T,w_{min}) \cdot D|)/2 \tag{4.26}$$

A score for documents could be given by the percentage of neighbouring term occurrences are separated by a distance in the desired interval, normalised by the total number of occurrences. The score would be:

% d in [0,10]		% d in [10,20]		% d in [90,100]	
predominant	36	the	18	and	9
red	36	of	17	in	9
tails	31	and	15	of	9
the	27	breadth	13	the	9
year	27	bud	13	to	8
monstrosities	24	flowers	13	a	7
began	22	generic	13	flowers	7
dissimilar	18	in	13	leaves	7
of	17	limit	13	limit	7
causes	15	pear	13	relative	7

Table 4.2: Terms with most inter-occurrence distances between 10 and 40 tokens in chapter 2 of "The Origin of Species"

$$S(T) = \frac{(2(w_{max} - w_{min})|E(T,0) \cdot D| + |E(T,w_{max}) \cdot D| - |E(T,w_{min}) \cdot D|)}{(length|E(T,0) \cdot D| - |E(T,(length-1)/2) \cdot D|)} \tag{4.27}$$

Scoring terms with this counting, for $w_{min} = 5, w_{max} = 20$ for the chapter of "The Origin of Species" used as an example in section 4.3.1. The highest scoring terms are shown in table 4.2

Another way of scoring terms within a document, is defining a critical width above which every token in the document is preserved, but below which something is erased: we will call it **Covering Width**:

Definition 4.9 (Covering Width)

Covering Width $w^*_{t,D}$ is the minimum value of w for which $E(t, w)$ will preserve the whole document D. This can only be defined if t occurs in D:

$$(|E(t, w^*_{t,D}) \cdot D| = |D|) \wedge (|E(t, (w^*_{t,D} - 1)) \cdot D| < |D|) \tag{4.28}$$

If a term appears uniformly spread through the sequence of the text, there will be little overlap, and the whole document will be covered by the windows around the central term with a relatively small width. Stopwords are expected to behave in this way. But if a term is frequent in the document, but its occurrences appear concentrated in part of the document, its covering width will be larger. This suggests another way of scoring terms within a document:

$$S(t \; in \; D) = \frac{2w^*_{t,D}}{|D|}|E(t, 0) \cdot D| \tag{4.29}$$

TERM	SCORE
selection	12.75
breeds	12.49
rockpigeon	6.59
flowers	6.18
plants	6.15
descended	5.90
wild	5.82
fruit	5.59
improved	5.14
savages	5.08

Table 4.3: Terms with best product [normalised term frequency]·[covering width] for chapter 2 of "The Origin of Species"

The procedure to obtain the score is much slower than choosing an interval of distances (specially

for a long document), but the results for the chapter of "The Origin of Species" shown in table 4.3 are intuitively much better than those of 4.2

4.4 Order Relations Between Erasers

When the parameter w is increased in a SE, the windows of preserved text will grow. Order relations are left unchanged when the windows of the including SE are expanded in the same amount or more than those in the included SE:

$$((E(a, w_{a,1}) \geqslant E(b, w_{b,1})) \wedge ((w_{a,2} - w_{a,1}) \geqslant (w_{b,2} - w_{b,1}) \geqslant 0) \Rightarrow (E(a, w_{a,2}) \geqslant E(b, w_{b,2})) \quad (4.30)$$

Since some order relations are trivially implied by others, it is useful to consider only those from which the other can be derived. Let us then define those that are **tight**, in the sense that they cannot be obtained from others. They can be defined as follows:

Definition 4.10 (Tight Inclusion)

Tight inclusion relation between two SEs: ($\succcurlyeq_C$) An inclusion relation that ceases to hold if the width of the including SE is decreased, or if the width of the included SE is increased:

$$(E(a, w_a) \succcurlyeq_C E(b, w_b)) \iff ((E(a, w_a) \geqslant_C E(b, w_b)) \wedge ((E(a, w_a - 1) \not\geqslant_C E(b, w_b)) \wedge (E(a, w_a) \not\geqslant_C E(b, w_b + 1)))) \quad (4.31)$$

Order relations provide a way of grouping SEs into **classes of equivalence**, that is, sets of SEs that are equivalent when applied to a set if documents:

Definition 4.11 (Equivalence Between Erasers)

Equivalence relation between two SEs: Two SEs are equivalent when they include each other.

$$((E(t_2, w_2) \leqslant_C E(t_1, w_1)) \wedge (E(t_2, w_2) \geqslant_C E(t_1, w_1))) \iff (E(t_2, w_2) \equiv_C E(t_1, w_1)) \quad (4.32)$$

An important property of a class of equivalence, is that if an element of one class of equivalence includes another, then any member of the first will include any member of the other:

$$\forall\,(E_{A1} \equiv_C E_{A2}, E_{B1} \equiv_C E_{B2})\,,$$
$$(E_{A1} \geqslant_C E_{B1}) \iff (E_{A1} \geqslant_C E_{B2}) \iff (E_{A2} \geqslant_C E_{B1}) \iff (E_{A2} \geqslant_C E_{B2}) \quad (4.33)$$

Two SEs in a class of equivalence will always preserve the same number of tokens, but the same number of tokens is not a sufficient criterion for equivalence:

$$(E_1 \equiv_C E_2) \Rightarrow (\forall D \in C, |E_1 \cdot D| = |E_2 \cdot D|) \quad (4.34)$$

Yet another possible relation between SEs is *disjointedness*.

Definition 4.12 (Disjointedness of Erasers)

Disjointness relation between two SEs: Two SEs are *disjoint* when, while neither of them erases everything in the document, their product does:

$$(E(t_1, w_1) \perp_C E(t_2, w_2))$$
$$\iff (\forall D \in C, (|E(t_2, w_2) \circ E(t_1, w_1) \cdot D| = 0) \wedge (|E(t_1, w_1) \cdot D| \cdot |E(t_2, w_2) \cdot D| > 0)) \quad (4.35)$$

The symbol $\perp_C$ has been chosen to represent disjointedness because this relation is very similar to the geometrical relation of orthogonality (also called perpendicularity). Three of its properties are:

- Anti-reflexive $\neg(E_1 \perp_C E_1)$
- Symmetric $(E_1 \perp_C E_2) \iff (E_2 \perp_C E_1)$
- Non-transitive $((E_1 \perp_C E_2) \wedge (E_2 \perp_C E_3)) \nRightarrow (E_1 \perp_C E_3)$

4.4.1 Necessary Order Relations Between Erasers

Some relations between SEs will hold for any imaginable text document. For example, a wider window SE centred in an occurrence of a term will always include another SE with narrower window centred on the same term

$$(w_1 \geqslant w_2) \Rightarrow (E(t, w_1) \geqslant E(t, w_2)) \tag{4.36}$$

Another important relation is that every pair of SEs with $w = 0$ centred on different terms, are disjoint.

$$\forall (D, t_1 \neq t_2), |E(t_1, 0) \circ E(t_2, 0) \cdot D| = 0 \tag{4.37}$$

4.4.2 Contingent Order Relations Between Erasers

Some of the order relations between erasers will only hold for a limited set of documents. We will call these relations *contingent*. Some of these will arise from the presence or absence of terms in the set of documents considered:

- Within a set of documents, all SEs centred in non-present terms will be equivalent, because all of them will simply erase all the text.

$$\forall D_i \in C, (|E(t_1, 0) \cdot D_i| = 0) \wedge (|E(t_2, 0) \cdot D_i| = 0) \\ \Rightarrow (\forall (w_1, w_2), E(t_1, w_1) \equiv_C E(t_2, w_2)) \tag{4.38}$$

- For terms that are present in *all* documents within a set, on the other hand, there will be a threshold $w^*_{t,C}$ above which all SEs will be equivalent, because they will preserve the whole text and erase no token. We are going to call this threshold **covering width for term t in set C**:

$$\forall D \in C, (|E(t,0) \cdot D| > 0)$$
$$\Rightarrow \left(\exists w^*_{t,C}, \forall w \geqslant w^*_{t,C}, |E(t,w) \cdot D_i| = |D|\right) \quad (4.39)$$

$$(|E(t_1,0) \cdot D| > 0) \wedge (|E(t_2,0) \cdot D| > 0)$$
$$\forall (w_1 \geqslant w^*_{t_1,C}, w_2 \geqslant w^*_{t_2,C}), E(t_1,w_1) \equiv_C E(t_2,w_2) \quad (4.40)$$

The most important kind of contingent relations is, however, that of inclusion of SEs centred in different terms that neither erase nor preserve the whole document. They are closely related to maximal and minimal distances between neighbouring term occurrences, and therefore are very sensitive to co-occurrence tendencies.

4.4.3 Occurrence Distances and Inclusion

Obtaining an expression for the distribution of distances between occurrences of two different terms is even simpler than obtaining one for distances between occurrences of the same term. The expression emerges in a trivial way when we consider that every occurrence of a given term T_1 in the windows defined by $E(T_2, w)$ (with $w > 0$) cannot be separated from an occurrence of T_2 by more than $w - 1$ tokens, that is, the maximum distance there could be between these preserved T_1 and T_2 is w:

$$N(d(t_1, closest\ t_2) \leqslant w) = |E(t_2,0) \circ E(t_1,w) \cdot D| \quad (4.41a)$$

$$N(d(t_2, closest\ t_1) \leqslant w) = |E(t_1,0) \circ E(t_2,w) \cdot D| \quad (4.41b)$$

These relations are, of course, related to inclusion relations, since, when one of two SE has $w = 0$ the following relation holds:

$$(|E(t_2,0) \circ E(t_1,w) \cdot D| = |E(t_2,0) \cdot D|) \Rightarrow (E(t_1,w) \geqslant_{\{D\}} E(t_2,0)) \quad (4.42)$$

The difference between the two distances defined in (4.41a) and (4.41b) can be subtle. To explain it, let us consider two terms that allow to produce a disjointed pair of SEs and two different couples of SEs with opposite including relations, like those represented in figure 4.8

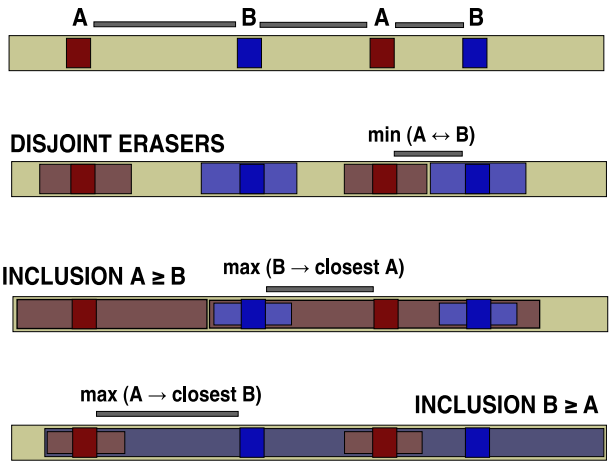

Figure 4.8: maximal and minimal distances determining the values of parameter w for which relations between SEs hold.

> Note that the maximum distances between an occurrence of A and the closest occurrence B is not the same maximum distance between an occurrence of B and the closest occurrence of A, because they can be different pairs of occurrences.

Let us suppose that terms t_1 and t_2 occur in every document in textual context C. If d_{min} is the minimum number of tokens between neighbour occurrences and $d_{max(t_1,t_2)}$ is the maximum number of tokens between any occurrence of t_1 and the nearest occurrence of t_2, we have already some nontrivial relations that hold this document:

$$E(t_1, w_1) \perp_D E(t_2, w_2) \iff (w_1 + w_2) \leqslant d_{min} \tag{4.43}$$

$$E(t_1, w_1) \geqslant E(t_2, w_2) \iff w_1 + 1 \geqslant (w_2 + d_{max}) \tag{4.44}$$

$$E(t_1, w_1) \leqslant E(t_2, w_2) \iff w_2 + 1 \geqslant (w_1 + d_{max}) \tag{4.45}$$

Distances between neighbouring occurrences of **different** terms are not totally independent from distances between re-occurrences of a **same** term, like those considered in section 4.3.1. Distances between A and B cannot be largeger than the maximum distance between a A and a border or another A, or than the maximum distance between a B and a border or another B.

For example: in sequence x x x A x x A x x B x x x A x x x B x x x x x x x x, no distance can be larger than 8, which is the maximal distance between one of the considered terms (B) and a border or another occurrence of itself. However, no lower bond is imposed on distances between occurrences of different terms: occurrences of A and B could appear together without being restricted by distances A-border, B-border, A-A or B-B.

The expected distribution of distances from A to the closest B in a totally random case can be obtained from that of A and B alone, by adding all the uniform distributions between 0 and a maximal distance determined by the individual pairs of distances between occurrences of A and B:

$$N(d_{A,B} \leqslant x) = \min(N_A + 1, N_B + 1) \sum_{x=1}^{length} \frac{1}{x} \left(\frac{N(d_A \leqslant x) N(d_B \leqslant x)}{(N_A + 1)(N_B + 1)} \right) \tag{4.46}$$

4.4.4 Eraser Lattices

Any set of equivalence classes of erasers, together with their order relations, constitute a *Partially Order Set* (poset). Relation $\geqslant$ fulfils the requirements of an order relation: it is reflexive, anti-symmetric and transitive. The set of equivalence classes of SEs is not *totally ordered* but *partially*, because it is not the case that every pair of Erasers are orderly related to each other in one way or the opposite ($\neg(\forall(E_1, E_2)\ \ (E_1 \geqslant E_2) \vee (E_2 \geqslant E_1))$) [92]. If every SE is considered, the poset is also a *lattice*, because there is a class of SEs that erase everything, and is therefore included by any other (called the *infimum* and there is a class of transformations that do not erase anything, and therefore includes any other (called the *supremum*).

These order relations can be defined on documents alone, or in sets of documents (textual contexts). Let us first examine how they look like for individual documents, and then for sets of documents.

4.4.5 Equivalence of SEs for a document

Equivalence relations can be a very simple way of representing information about the usage of terms beyond bag-of-words. All erasers $E(t, w)$ will be equivalent whenever $w \geqslant w^*_{t,D}$. As widths

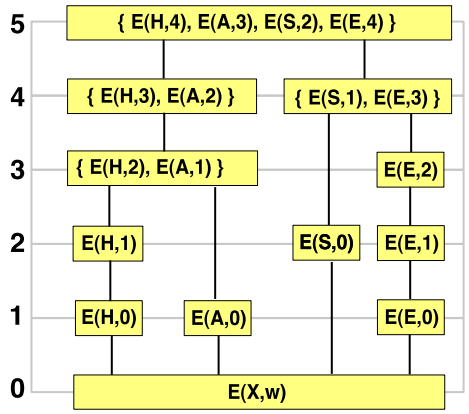

Figure 4.9: Hasse diagram for the similarity classes of letter SE as they behave on the word "H A S S E"

Horizontal rows correspond to the preserved token (letter) count, and vertical lines correspond to order relations. Yellow boxes are similarity classes of letter SEs, defined by their action on the word. The lower class corresponds to all the SEs centred on letters not appearing on HASSE, and the upper class will include the enumerated SEs, plus any other resulting by increasing factor w to any of those.

decrease from that value, some SEs centred in different terms cease to be equivalent.

The number of classes of equivalence for a single given document can be bounded in a simple way. If a document contains N_D different terms, then it should at least have the classes of equivalence corresponding to the 0-width SEs $E(t, 0)$, plus supremum and infimum; that is, $N_D + 2$ classes of equivalence. The maximum number of classes that it could have, on the other hand, is limited not only by the number of distinct terms, but also by its length; for any term t present in the document, any $E(t, w)$ will be in the equivalence class of the supremum if $2w + 1 \geqslant length$. This means:

$$N_D + 2 \leqslant N_{classes} \leqslant length \cdot N_D + 2 \tag{4.47}$$

This, together with the fact that every element in a SE class of equivalence will produce the same preserved token count (shown in (4.34)) suggests a representation for lattices in a 2-dimensional array, as follows:

- Each central term is assigned a column
- Each possible token counting (from 0 to $length$) is assigned a row

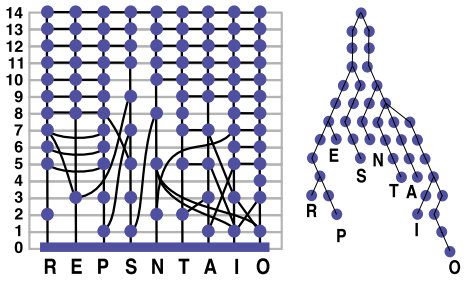

Figure 4.10: Diagrams representing SEs centred on different letters of the word "R E P R E S E N T A T I O N".

On the left, a binary array with rows representing the number of preserved letters, and columns representing the central letter, with order relations represented as lines (horizontal lines are equivalence relations, in order relations the higher point includes the lower). On the right, a tree showing the classes of equivalence. They start being one for the covering width and split in several as width decreases.

- SEs are represented by dots (or circles) in the appropriate sites of the array
- Lines between the dots represent order relations. If the line is horizontal, it represents equivalence, otherwise, it means that the upper includes the lower.
- A rectangle in the lower row (0 preserved tokens) represents the class of SEs centred on absent terms.

This can be thought as a more elaborate version of the lattice represented in figure 4.9. A more complex one is shown in figure 4.11, this one with letter-erasers for the word "REPRESENTATION".

Some features of the usage of terms in the document can be easily seen in the lattice representation, for example:

- The row where lowest-lying dot in a column is, will show the occurrence counting of the corresponding term $|E(t,0) \cdot D|$.
- If occurrences of the term are evenly distributed in the document sequence, the dots on its corresponding column will have gaps between them. If, on the contrary, the occurrences appear mostly in a small part of the document, the dots will be closer together, and there will be more of them.

- If two terms tend to co-occur, the corresponding columns will appear as connected by a large number of lines.

- Equivalence classes of SEs will indicate parts of the document where terms in a subset occur together. Every SE in an equivalence class will preserve the same portion of text, and this portion will contain the occurrences of the central terms corresponding to other SEs in the class within it. The token count of the class (the row where they are) will indicate the size of these clusters, that could also be composed of several disconnected areas along the document.

4.4.6 Representing the Lattice

Figure 4.11 suggests a way to represent eraser token counting in a document with a sparse binary array:

Definition 4.13 (Erasers Array)

Erasers Array of a document: with a column index corresponding to a term, and a row index corresponding to a token counting that can vary from 1 to $|D| - 1$. Its entries are binary: 0 if no SE centred in the term produces the corresponding token counting, and 1 if it does.

The value of $w*_{t,D}$ would be simply the number of 1s in the column corresponding to t. Erasers for the chapter of "The Origin of Species" are represented in figure 4.11 for the 500 most frequent terms.

The erasers array is very easy to approximate by relations like 4.24, or an even simpler scheme, like simply adjusting the relation between $|E(t,w) \cdot D| = f(|E(t,0) \cdot D|, w)$ by only measuring, for example, the width for which half of the text is preserved, plus the width for which all of the text is preserved (covering width). This amounts to assuming a form for the distribution of distances, as was shown in section 4.3.2

As for the order relations between SEs centred on different terms, equivalence relations capture most of the information for large fractions of preserved text (as can be seen in figure). All SEs preserving the whole document will belong to the same equivalence class, but this equivalence

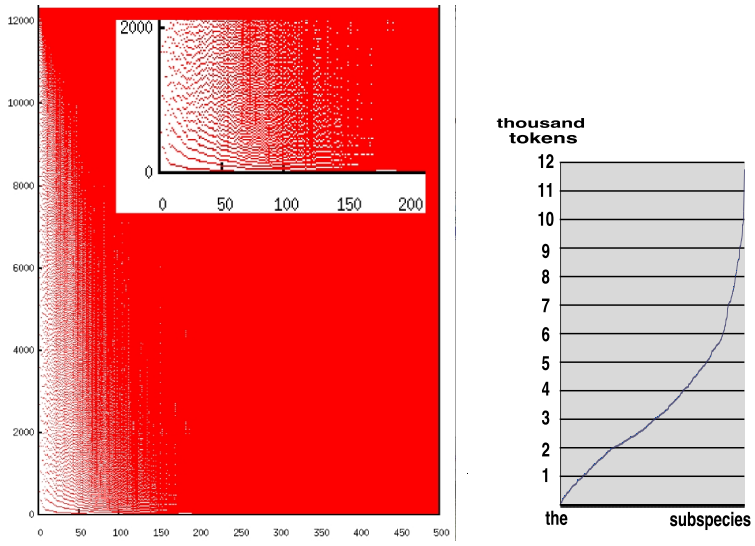

Figure 4.11: Left: Binary Array representing the useful SEs in chapter 2 of "The Origin of Species" centred on 500 of the most frequent terms. Right: Covering widths for terms, from more frequent to less frequent (is also the density of each column in the graphic on the left)

> In the graph on the left, each column (axis x corresponds the set of 500 most frequent terms, ordered from most frequent to less frequent ("the" appears 828 times, while the last term, "subspecies" appears only 4 times). Each row corresponds to a token counting, and goes from 4 (only 4 token preserved) to 12323 (all the tokens in the whole text). Because of the scale, it is hard to appreciate the density of points; in the right graph the number of points for each column (which corresponds to the *covering width*) is shown for each term.

class splits progressively as the number of preserved tokens decreases, and classes are smaller and smaller, until they consist of only 1 element for small fractions of preserved text.

4.4.7 Eraser Lattices for a Set of Documents

The conditions for any order relation between SEs are more strict in a set of documents than they are in a single one, because they have to be fulfilled for all of them. This means that the overall number of order relations will decrease. Relations involving token counting for a set of documents will still hold, provided that the sum of counts over all documents in the set replace single document counts:

Inclusion $\geqslant_C$:

$$(\sum_{D\in C} |E(t_2,0) \circ E(t_1,w) \cdot D| = \sum_{D\in C} |E(t_2,0) \cdot D|) \;\Rightarrow\; (E(t_1,w) \geqslant_C E(t_2,0)) \qquad (4.48a)$$

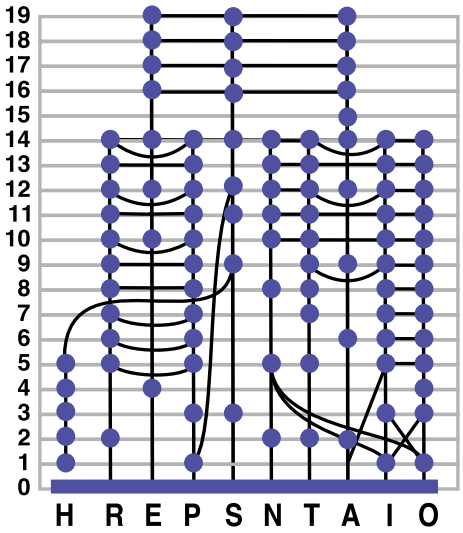

Figure 4.12: Lattice of term SEs for two words: "H A S S E" and "R E P R E S E N T A T I O N"

Only letters common to both, (A, S and E) can be in the top row, as part of the *supremum* equivalence class

Equivalence $\equiv_C$:

$$(E_1 \equiv_C E_2) \Rightarrow (\sum_{D \in C} |E_1 \cdot D| = \sum_{D \in C} |E_2 \cdot D) \tag{4.48b}$$

Covering width $w^*_{t,C}$:

$$\forall t | \prod_{D \in C} |E(t,0) \cdot D| > 0, \exists w^*_{t,C}, \sum_{D \in C} |E(t, w^*_{t,C}) \cdot D| = \sum_{D \in C} |D| \tag{4.48c}$$

The disappearance of some order relations when adding documents to the set will be reflected also in the splitting of the equivalence classes: there are more of them, and smaller, for a set of documents than for one alone.

There is a further qualitative difference with the case of one single document: a SE preserving all text in the set of documents can only have a central term occurring in every document in the set. In the graphic, this means that not all columns will have dots in the upper row, corresponding to the sum of lengths.

All these effects can be seen in figure 4.12

4.5 Representation of Documents with Lexical Measurements

Bag of words representations are very compact, but they capture semantic content in a remarkable way. A document represented by the set of frequencies $\{N_t\}$ can represent a very large number of possible scrambled documents. In fact, the number of possible text sequences represented by a vector of term frequencies is:

$$N_{sequences} = \frac{(\sum_t N_t)!}{\prod_t (N_t!)} \tag{4.49}$$

To take an example, document AP880121-001 from TREC collections, has 249 distinct terms and is a sequence of 383 tokens. The distribution of numbers of occurrences is as follows:

N_t	25	15	8	7	6	5	4	3	2	1
terms with N_t	1	1	2	1	3	6	5	9	38	149

Applying this, we find that the TF vector for this documents would be the same for a huge number of possible sequences:

$$N_{sequences} = \frac{383!}{(25!)(15!)(8!)^2(7!)(6!)^3(5!)^3(4!)^5(3!)^9(2!)^{38}} \approx 4.47 \times 10^{738} \tag{4.50}$$

Most of these sequences will neither make any sense, nor comply with syntax rules that exist in natural language. It is, indeed, extremely difficult to compute the fraction of sequences that make any sense, but intuitively we can think of some permutations of words that will still produce a text dealing with the same topic, and others that will produce a text dealing about something else. For example, permutations of synonym words will probably not affect semantic contents much, and permutations of terms surrounding a preposition is likely to change the meaning of the sentences ("X of the Y" is likely to have a different meaning than "Y of the X").

4.5.1 Comparison of Documents

The operations and relations between SEs introduced so far allow us to formally define an operative notion of identity between documents:

Definition 4.14 (Equality of Erasers)

Two documents are measured **equal** when every product of SEs preserves the same number of tokens in both.

$$(D_1 = D_2) \iff \forall\{(t_1, w_1), (t_2, w_2), \cdots, (t_N, w_N)\}|(\prod_{i=0}^{N} E(t_i, w_i))\cdot D_1| = |(\prod_{i=0}^{N} E(t_i, w_i))\cdot D_2| \tag{4.51}$$

For most products of SEs, the number of preserved tokens will be 0, but there is an interesting set for which this number is different. For a given document, it is possible to define a set of *optimally discriminating products*. These products will impose the most rigid restrictions to the preserved text, while still preserving a number of tokens larger than 0.

Definition 4.15 (Optimally Discriminating Product)

An **Optimally preserving product** is a product of SE that fulfills the following conditions:

1. Includes the highest possible number of SEs with different central terms.
2. The sum of the widths is minimal.

It is possible also to define optimally discriminating products with a fixed number of SEs. In this case, condition 1 of the definition would be changed.

As an example of maximally discriminating products, let us consider the sentence D_1 ="**To be or not to be, that is the question**". This sentence has 8 terms and 10 tokens. 10 products P_i containing SEs centred in every term present in the sentences can be built to preserve one token in

different places of the text sequence:

$$
\begin{aligned}
P_1 &= E(\mathbf{to},\mathbf{0})E(\mathbf{be},\mathbf{1})E(\mathbf{or},\mathbf{2})E(\mathbf{not},\mathbf{3})E(\mathbf{that},\mathbf{6})E(\mathbf{is},\mathbf{7})E(\mathbf{the},\mathbf{8})E(\mathbf{question},\mathbf{9})\\
P_2 &= E(\mathbf{be},\mathbf{0})E(\mathbf{or},\mathbf{1})E(\mathbf{to},\mathbf{2})E(\mathbf{not},\mathbf{3})E(\mathbf{that},\mathbf{6})E(\mathbf{is},\mathbf{7})E(\mathbf{the},\mathbf{8})E(\mathbf{question},\mathbf{9})\\
P_3 &= E(\mathbf{or},\mathbf{0})E(\mathbf{not},\mathbf{1})E(\mathbf{be},\mathbf{2})E(\mathbf{to},\mathbf{3})E(\mathbf{that},\mathbf{6})E(\mathbf{is},\mathbf{7})E(\mathbf{the},\mathbf{8})E(\mathbf{question},\mathbf{9})\\
P_4 &= E(\mathbf{not},\mathbf{0})E(\mathbf{to},\mathbf{1})E(\mathbf{or},\mathbf{2})E(\mathbf{be},\mathbf{3})E(\mathbf{that},\mathbf{5})E(\mathbf{is},\mathbf{7})E(\mathbf{the},\mathbf{8})E(\mathbf{question},\mathbf{9})\\
P_5 &= E(\mathbf{to},\mathbf{0})E(\mathbf{be},\mathbf{1})E(\mathbf{not},\mathbf{2})E(\mathbf{that},\mathbf{3})E(\mathbf{or},\mathbf{4})E(\mathbf{is},\mathbf{5})E(\mathbf{the},\mathbf{7})E(\mathbf{question},\mathbf{9})\\
P_6 &= E(\mathbf{be},\mathbf{0})E(\mathbf{that},\mathbf{1})E(\mathbf{to},\mathbf{2})E(\mathbf{is},\mathbf{3})E(\mathbf{not},\mathbf{4})E(\mathbf{the},\mathbf{5})E(\mathbf{or},\mathbf{6})E(\mathbf{question},\mathbf{7})\\
P_7 &= E(\mathbf{that},\mathbf{0})E(\mathbf{is},\mathbf{1})E(\mathbf{be},\mathbf{2})E(\mathbf{the},\mathbf{3})E(\mathbf{to},\mathbf{4})E(\mathbf{question},\mathbf{5})E(\mathbf{not},\mathbf{6})E(\mathbf{or},\mathbf{7})\\
P_8 &= E(\mathbf{is},\mathbf{0})E(\mathbf{the},\mathbf{1})E(\mathbf{that},\mathbf{2})E(\mathbf{question},\mathbf{3})E(\mathbf{be},\mathbf{4})E(\mathbf{to},\mathbf{5})E(\mathbf{not},\mathbf{6})E(\mathbf{or},\mathbf{7})\\
P_9 &= E(\mathbf{the},\mathbf{0})E(\mathbf{question},\mathbf{1})E(\mathbf{is},\mathbf{2})E(\mathbf{that},\mathbf{3})E(\mathbf{be},\mathbf{4})E(\mathbf{to},\mathbf{5})E(\mathbf{not},\mathbf{6})E(\mathbf{or},\mathbf{7})\\
P_{10} &= E(\mathbf{question},\mathbf{0})E(\mathbf{the},\mathbf{1})E(\mathbf{is},\mathbf{2})E(\mathbf{that},\mathbf{3})E(\mathbf{be},\mathbf{4})E(\mathbf{to},\mathbf{5})E(\mathbf{not},\mathbf{6})E(\mathbf{or},\mathbf{7})
\end{aligned}
$$

If these products are applied to a sentence that is very similar in lexical terms but has a different meaning, $D_2 =$ "**The question is, to be or not to be that**", the counts of preserved tokens would be all null, because some of the factors would be orthogonal for this permuted sentence:

$$
\begin{array}{lll}
|P_1D_1| = 1 & |P_1D_2| = 0 & |E(that,6) \circ E(the,8) \cdot D_2| = 0\\
|P_2D_1| = 1 & |P_2D_2| = 0 & |E(that,6) \circ E(the,8) \cdot D_2| = 0\\
|P_3D_1| = 1 & |P_3D_2| = 0 & |E(that,6) \circ E(the,8) \cdot D_2| = 0\\
|P_4D_1| = 1 & |P_4D_2| = 0 & |E(that,5) \circ E(the,8) \cdot D_2| = 0\\
|P_5D_1| = 1 & |P_5D_2| = 0 & |E(that,3) \circ E(the,7) \cdot D_2| = 0\\
|P_6D_1| = 1 & |P_6D_2| = 0 & |E(that,1) \circ E(the,5) \cdot D_2| = 0\\
|P_7D_1| = 1 & |P_7D_2| = 0 & |E(that,0) \circ E(is,1) \cdot D_2| = 0\\
|P_8D_1| = 1 & |P_8D_2| = 0 & |E(is,6) \circ E(the,1) \cdot D_2| = 0\\
|P_9D_1| = 1 & |P_9D_2| = 0 & |E(question,1) \circ E(that,3) \cdot D_2| = 0\\
|P_{10}D_1| = 1 & |P_{10}D_2| = 0 & |E(question,0) \circ E(that,3) \cdot D_2| = 0
\end{array}
$$

Neither the problem of finding maximally discriminating products, nor the uniqueness of a set of them, are solved in this work. However, a simple approximate procedure to produce a set of such products for a particular document D is proposed in appendix D.

4.5.2 Vector-Space Similarity between Documents, and SEs

Representing a document by a vector containing term frequencies, or some function of them, is one of the oldest ideas in IR that still remains fruitful. As was shown in section 2.3.2, a document

is represented by a vector in a space whose natural basis consists in a set or orthonormal vectors corresponding to a term each.

With a weighted norm scheme like that defined in 4.2.2, it is possible to reproduce precisely the similarity measure as it is developed in vector space models (see section 2.3.2), just by estimating some of the representation coefficients:

$$\langle t_i|D\rangle = |E(t_i,0)\cdot D| \qquad |\lambda_j\rangle = |t_j\rangle \tag{4.52}$$

Similarity between documents would be computed as it is shown in 2.8. Including SEs the relation would be:

$$\vec{D}_1\cdot\vec{D}_2 = \sum_{t_i}|E(t_i,0)D_1|\times|E(t_i,0)D_2| \tag{4.53}$$

where $\vec{D}$ is the vector bag-of-words representation of a document. This is precisely part of a known measure of similarity: the *cosine similarity* [46]. In section 2.3.2 we mentioned that in Vector Space Models terms are considered as elements of a dual space to documents, so that a numeric coefficient corresponds to every term-document pair. In this new definition of inner product with SEs, an eraser corresponds to every term, forming an **operator valued measure**, or more explicitly an **eraser valued measure**. This is analogous to the concept of *Positive-defined Operator-Valued Measure* (POVM) [117]

The equivalence classes tree shown in figure 4.11 gives some information that can be approximately encoded in differences between the phases of entries corresponding to different terms. From the tree, we can define distances between terms, and assign phases whose differences reflect distances between terms in the tree.

$$\langle A|D\rangle\langle D|B\rangle \approx \left(\frac{\sqrt{N_{A\text{ in }D}N_{B\text{ in }D}}}{length}\right)e^{i\pi\left(\frac{d_{A,B}}{d_{max}}\right)} \tag{4.54}$$

where $i=\sqrt{-1}$, $d_{A,B}$ is the distance between terms in the tree, and $N_{X\text{ in }D}$ is the frequency of term X in document D. If we take terms as an orthonormal basis $\{|t_j\rangle\}$, and distances as anti-symmetric quantities $d_{A,B}=-d_{B,A}$, equation (4.54) gives us the expected entries of the matrix representing the rank-one projector $|D\rangle\langle D|$ in this basis.

$$\langle t_j|D\rangle\langle D|t_k\rangle \approx M_{j,k} = \left(\frac{\sqrt{N_{t_j \text{ in } D} N_{t_k \text{ in } D}}}{length}\right) e^{2\pi i\left(\frac{d_{t_j,t_k}}{d_{max}}\right)} \tag{4.55}$$

The resulting matrix will not be, in general, a rank-one projector, but will have the mathematical characteristics of a density operator (normalised, positive-defined). Two ways of defining similarity between density operators are *fidelity* and *normalised frobenius product*:

Definition 4.16 (fidelity)

Fidelity measures how the probabilities of any measurement given a state represented by a density operator, are reproduced by the probabilities of the same measurements by a state represented by another density operator [118]. The mathematical definition is:

$$F(\rho_1, \rho_2) = \text{Trace}\left((\rho_1)^{\frac{1}{4}}(\rho_2)^{\frac{1}{2}}(\rho_1)^{\frac{1}{4}}\right) \tag{4.56}$$

Fidelity seems to be a very natural choice to compare density matrices, but calculating it can be computationally demanding. An alternative is a normalised frobenius product, which is equivalent when density operators are rank-one projectors:

Definition 4.17 (Frobenius Normalised Product)

Frobenius Normalised Product is a measure of similarity between matrices that is defined as a generalisation of the cosine product for vectors. The formal definition is:

$$FNP(\rho_1, \rho_2) = \sqrt[4]{\frac{\text{Trace}(\rho_1(\rho_2)^2\rho_1)}{\text{Trace}\left((\rho_1)^2\right)\text{Trace}\left((\rho_2)^2\right)}} \tag{4.57}$$

If a vector representation of documents is needed (for example, for LSI), it can also be derived in an approximate way from the matrix M. Vector $|D\rangle$ can be computed by maximising $\langle D|M|D\rangle$; this amounts to taking the eigenvector of M with the highest eigenvalue. This is a way of producing a basic vector space approach including information beyond the bag-of-words approach.

4.6 Non-Boolean Algebra on Erasers

In the definitions of the Boolean operations (4.3, 4.2, and 4.4) it is implied that the central terms to define each SE are "visible", even if they have been erased by another SE before. If the analogy

with quantum measurement is to be followed, however, the possibility that they are overlooked has to be considered. For this, it is necessary to use successive application (products) of SEs.

Definition 4.18 (Quantum Meet (intersection))

q-meet of two SEs $E(t_1, w_1) \wedge_q E(t_2, w_2)$: Is the transformation that erases the tokens that are erased by the successive application of both SE in any order:

$$E(t_1, w_1) \wedge_q E(t_2, w_2) \equiv (E(t_1, w_1) \circ E(t_2, w_2)) \wedge (E(t_2, w_2) \circ E(t_1, w_1)) \tag{4.58}$$

A q-join cannot be defined in a similar manner, as $E(t_1, w_1) \vee_{q^*} E(t_2, w_2) \equiv (E(t_1, w_1) \circ E(t_2, w_2)) \vee (E(t_2, w_2) \circ E(t_1, w_1))$ because this could preserve less tokens than each of the SEs, producing therefore monstrous order relations. It is desirable to define a q-join that is just as prone to preserve tokens as the q-meet is to erase them, producing therefore an algebra whose relations are in a way equally valid when dealing with the negation of every element. This symmetry of the algebra can be formulated as the de Morgan's law:

Definition 4.19 (De Morgan's Law)

De Morgan's law :

$$A \wedge B = \neg(\neg A \vee \neg B) \tag{4.59}$$

The de Morgan's law can be used to define a q-join that preserves this symmetric character of the algebra:

Definition 4.20 (Quantum Join (Union))

q-join of two SEs $E(t_1, w_1) \vee_q E(t_2, w_2)$: Is the transformation that fulfils the de Morgan's Law with the q-meet:

$$E(t_1, w_1) \vee_q E(t_2, w_2) \equiv \neg(\neg E(t_1, w_1) \wedge_q \neg E(t_2, w_2)) \tag{4.60}$$

Finite Quantum Logics are known to fulfil de Morgan's law [119], so this definition seems appropriate to continue the parallel with the quantum-theoretical notions of measurement.

4.6.1 Is a Logic of Erasers distributive?

As was mentioned in section 3.4, a key property of Quantum Logics is that distributive law is not universally fulfilled. It can be shown that this law can also be violated by the algebra $\{E(t_i, w_i), \neg, \wedge_q, \vee_q\}$. One example is enough for that:

Example 3

Suppose the document is a sequence of terms:

D = `A xxB xxC xA B xxB xC xxA xxx`.

We define $E(A,2)$, $E(B,2)$ and $E(C,3)$, and check wether the following equality holds:

$$E(A,2)\wedge_q(E(B,2)\vee_q E(C,3))\cdot D =^{?} (E(A,2)\wedge_q E(B,2))\vee_q(E(A,2)\wedge_q E(C,3))\cdot D \quad (4.61)$$

To check this, we translate it to products and Boolean operations $\wedge$. This results in a rather long expression.

On the left side, everything boils down to the meet of products of three transformations on the left side (involving $E(A,2)$, $\neg E(B,2)$ and $\neg E(C,3)$). On the right side, there are products of complements of products, involving $\neg(E(A,2)E(B,2))$, $\neg(E(A,2)E(C,3))$, and the same products in opposite order.

In the end, the left-hand side transformation turns out to preserve less of the sequence than the right-hand side

$$T_{left\ side}D = \texttt{A xx- - - - - - - - - - - - - - - - xx-}$$
$$T_{right\ side}D = \texttt{A xxB xxC - - - - - - xC - - - xxx}$$

4.7 Erasers and Probabilities

Up to this point, we have considered distribution of probabilities that different terms occupy a certain position in the sequence of a text document. This allows to quantify the information contained before and after erasure. However, probabilities defined in this way are not too useful. In this section, we will consider probabilities that are more close to the actual reading of documents, even though they will be considered from an abstract point of view.

In section 4.2.2 it was shown that it is possible to define a measure for documents with SEs. To define probabilistic spaces, it is necessary to impose more conditions on this measure (these are explained in depth in [120]). To move in that direction, we will now, instead of defining a measure for documents with SEs, use documents to provide a measure for SEs. Having defined a join and a meet, and shown that SEs and the transformations formed with them constitute a lattice, we can enunciate the requirements of a probability measure for SEs.

Definition 4.21 (Probability Measure)

A **Probability Measure** for SEs would be a function that assigns a real number between 0 and 1 to every SE, $M : E \to [0, 1]$, such that:

1. $M(E_1 \wedge_q E_2) \leqslant M(E_1)$
2. $M(E_1 \wedge_q E_2) \leqslant M(E_2)$
3. $(M(E_1 \wedge_q E_2) = 0) \Rightarrow (M(E_1 \vee_q E_2) = M(E_1) + M(E_2))$
4. $M(\neg E) = 1 - M(E)$

It can be proven that the **counting of preserved tokens of a document that has been transformed by the SE is indeed a probability measure, when appropriately normalised by the token counting of the document**. An intuitive way of seeing this probability measure in terms of a random process is as the answer to the following question: If we pick a token at random (with no bias) from a document, what is the probability that the considered SE will preserve this token?

$$M_D(E) \equiv \frac{|E \cdot D|}{|D|} \tag{4.62}$$

A probability distribution of documents $\rho_{\{D_i,P_i\}}$ (where $\sum_i P_i = 1$) can also be a probability measure for SEs, if defined as follows:

$$M_{\rho_{\{D_i,P_i\}}}(E) \equiv \sum_i P_i \frac{|E \cdot D_i|}{|D_i|} \tag{4.63}$$

This bears an obvious resemblance to density operators, which can be thought of as a probability distribution of vectors. Density operators constitute probability measures for projectors, as is stated by Gleason's Theorem [121]. The correspondence between documents and density operators completes then the formal parallel between the Quantum Theoretical representation of ideal measurement outcomes (projectors) plus system states (density operators), and SEs as lexical measurements plus documents (or distributions of them).

This can sound very theoretical and abstract, but particular formulations of these eraser-document probabilities correspond to concepts that are already very useful in different Natural Language Processing techniques, like Term Frequency, Document Frequency, and Co-Occurrence probability:

- Term Frequency. This can be defined with a null-width SE $E(t,0)$:

$$TF(t_i) = P(\text{token preserved in document } D \text{ by eraser} E(t_i,0)) = \frac{|E(t,0) \cdot D|}{|D|} \tag{4.64}$$

  where $TF(t_i)$ represents the frequency of term t_i, normalised with the total number of terms in the document.

- Document Frequency. This can be defined with very wide SE:

$$DF(t) = \sum_i P(\text{token preserved in document } D_i \text{ by eraser} E(t,w)) = \\ = \sum_i \frac{|E(t,w) \cdot D_i|}{|D_i|} \tag{4.65}$$

  where $DF(t_i)$ is the number of documents where the term occurrs. Width factor w must equal or surpass the maximum length of documents in the considered collection, to ensure the preservation of the whole text when the central term is present.

- Co-occurrence probability. This can be defined using products of Erasers:

$$P(t_1 \text{closer than } w \text{ to } t_2 \text{ in } D) = \frac{|E(t_2,0) \circ E(t_1,w) \cdot D|}{|D|} \tag{4.66}$$

or as a conditional probability:

$$P(t_1|t_2 \text{ is at a distance of } w \text{ or less}) = \frac{|E(t_1,0)E(t_2,w)D|}{|E(t_2,w)D|} \tag{4.67}$$

If the positions of tokens in a document are considered as elements in a set, the action of a SE on this document can be seen splitting this set in the set of preserved positions and the set of erased positions. These sets can be combined by the usual set operations, and the propositions that define them with the corresponding Boolean logical operations: the union (join, conjunction) intersection (meet, disjunction) or complement (negation).

4.8 A Linear Algebra for Erasers

Equation (4.63) for a probabilistic measure defined on SEs with a distribution of documents, can be expressed in a more elegant way if we are able to put the SE apart from the rest of the factors as follows:

$$P_{\{P_i,D_i\}}(E) = \sum_i P_i \frac{E \cdot D_i}{|D_i|} = |E \left(\sum_i \frac{P_i D_i}{|D_i|} \right)| = |E \cdot \rho_{\{P_i,D_i\}}| \tag{4.68}$$

This is a more formal way of obtaining the positive-defined, hermitian density operator ρ_{P_i,D_i} that was mentioned in section 4.5.1.

To be able to define operators like that in equation (4.68), we need some operations, like these:

1. Multiplication by a scalar

$$|E \cdot (\alpha D)| = \alpha |E \cdot D| \qquad\qquad |(\alpha E) \cdot D| = \alpha |E \cdot D| \tag{4.69}$$

2. Sum

$$|E(D_1 + D_2)| = |ED_1| + |ED_2| \qquad |(E_1 + E_2)D| = |E_1D| + |E_2D| \tag{4.70}$$

Note that the sum of SEs is equivalent to the join, provided that they do not overlap.

Using the mathematical properties of SE as operators, it is possible to derive from them composite operators in a number of ways. In the next chapter, we will use these mathematical tools to build a transformation that can be engineered to be sensitive to high-level attributes of the text, like the topic that it is about.

4.8.1 Term Co-Occurrences and Kernels from SEs

Limiting the description of texts to frequency of occurrences of terms has been a useful but largely criticised approach; and to avoid it was a main motivation for the **Generalised Vector Space** model proposed by Wong et al in 1985 [46]. The use of term co-occurrence in vector space models has been shown to be an example of a concept that is known in Machine Learning as the *Kernel trick* [47]. It consists in using a set of features to define a similarity between objects. This trick avoids some of the problems that complex dependencies amongst the features may cause.

4.8.2 Linear Algebra, and something more on Kernels

The features of a vector-space representation can sometimes be redundant or ill-scaled, leading to inefficiency and error-prone results; an optimal representation should be unbiased and made of features that perfectly complement each other. If information is available on a large collection, a representation that is unbiased and as uniform as possible for the particular collection can be built. The first step is to find a way of converting the average density matrix representation into the maximally uninformative one:

$$\langle j|\rho_{\text{any document in C}}|k\rangle = \frac{1}{N_D}\delta_{j,k} \tag{4.71}$$

This means that the representation of the least informative density operator in an optimal vector space would be a normalised density matrix[1]. The relation of the raw term representation to the optimal representation can be obtained from the relation of the uninformative density operator and the identity:

$$\frac{1}{N_D}\mathbb{I} = M\rho_{\text{any document in c}}M \tag{4.72}$$

where M is a coordinate transformation. With a spectral decomposition of the density operator, this transformation can be easily computed

$$\rho_{\text{any document in C}} = \sum_{i=1}^{Rank} P_i|\psi_i\rangle\langle\psi_i|$$

$$M = \sum_{i=1}^{Rank}\sqrt{\frac{1}{P_i}}|i\rangle\langle\psi_i| \tag{4.73}$$

The term-term kernel for the collection would then be:

$$K_{t_1,t_2} = (M^\dagger M)_{t_1,t_2} = \sum_{i=1}^{rank}\frac{1}{P_i}\langle t_1|\psi_i\rangle\langle\psi_i|t_2\rangle \tag{4.74}$$

In this optimal representation the average document density operator is forced to be non-informative: a normalised identity with maximum von Neumann entropy.

The transformation M that turns into a normalised identity can be interpreted as latent features constituted as term combinations, while terms themselves are not represented by orthogonal or normalised basis anymore, but by vectors whose product is given by the kernel:

$$\langle t_1|t_2\rangle = K_{t_1,t_2} \tag{4.75}$$

SEs can then be used in a number of ways to compute kernels; their use can give way to a number of mathematical techniques that start from a bag-of-words representation of documents, or from more complex approaches. However, to exploit the full power of the proposed scheme, more of the rich formal structure of relations and properties of SE is needed, like the definition of logic (or

[1] A normalised identity is the density operator with a maximum von Neumann entropy, so it is, in fact, the most uninformative density operator.

otherwise) binary operations, that are usually referred as constituting elements of an algebra [12]

4.9 Uncertain Conditional and Quantum Representations

One of the most suggestive ideas in IR has been to estimate aboutness as the probability of a logical relation: the document implying the query [8]. As it was mentioned in 3.5, the idea has led to most logic-based models in IR [122], and was also the first call for methods beyond Boolean logics, in particular, using quantum-like logics [10]. Representing both documents and queries as states where relations between SEs hold or not leads quite naturally not only to establish logical relations between them, but also to assess the probability of a logical relation given an incomplete knowledge.

In this work, the problem of assessing the degree to which a document implies a query is approached as the problem of how much **lexical features of the document imply lexical features of the query**. A quantum-like projector logic provides both a way to understand that question and a natural way to answer it,

4.9.1 A Vector Space for Erasers

Representing a set of SEs is, unsurprisingly, a bit more complicated than simply representing terms in a bag-of-words approach. Instead of defining a set of vectors, each corresponding to a term, a set of orthogonal basis sets must be defined. To represent the SEs centred on each term, an orthogonal basis $\{|t_j, k\rangle\}$ has to be chosen, and projectors $\Pi_{(t,w)}$ to represent SEs in the space spanned by these basis are defined the following way:

$$E(t_j, k) \rightarrow \Pi_{(t_j,k)} = \sum_{l=1}^{2k+1} |t_j, l\rangle\langle t_j, l| \tag{4.76}$$

With this formula, a projector representing $E(t_j, k)$ will have a rank of $2k + 1$. It is easy to verify that necessary order relations hold:

$$\forall(k \geqslant l), \left(\Pi_{(t_j,k)} = \Pi_{(t_j,l)} + \underbrace{\sum_{m=2l+2}^{2k+1} |t_j, m\rangle\langle t_j, m|}_{\text{orthogonal to } \Pi_{(t_j,l)}} \right) \Rightarrow \left(\Pi_{(t_j,l)}\Pi_{(t_j,k)} = \Pi_{(t_j,l)}\right) \tag{4.77}$$

In vector space models, a starting point is usually a non-collection-dependent representation of terms: for example, as members of an orthogonal basis (basic vector space) or as random vectors in a high-dimensional space (Random Projection [123]). Then, information about the collection can be used to generate a better representation (usually as a non-orthogonal set of vectors, where the mutual overlap reflects statistical correlation in their occurrence). A starting point for representing SEs would be a set where no relations hold between projectors representing SEs centred on different terms, as is expected in a large corpus.

This can be achieved by imposing a mutual overlap condition to the basis sets representing vectors:

Definition 4.22 (Mutual Overlap)
Two basis sets $\{|t_j, k\rangle\}$ and $\{|t_l, m\rangle\}$ present **mutual overlap** when no member of one is orthogonal to a member of the other:

$$\forall(k, m), |\langle t_j, k|t_l, m\rangle|^2 > 0 \tag{4.78}$$

Two sets of projectors $\{\Pi_{(a,j)}\}$ and $\{\Pi_{(b,k)}\}$ built with mutually overlapping basis will not present any crossed order relation like $\Pi_{(a,j)} \geqslant \Pi_{(b,k)}$. However, order relations relative to one vector can exist. Primitive inclusion relations relative to a vector would be defined as follows:

$$\begin{aligned}\left(\Pi_{(a,j)} \succcurlyeq_{|\psi\rangle} \Pi_{(b,k)}\right) \iff & \\ & (\langle\psi|\Pi_{(a,j)}\Pi_{(b,k)}\Pi_{(a,j)}|\psi\rangle = \langle\psi|\Pi_{(b,k)}|\psi\rangle) \wedge \\ & (\langle\psi|\Pi_{(a,j-1)}\Pi_{(b,k)}\Pi_{(a,j-1)}|\psi\rangle < \langle\psi|\Pi_{(b,k)}|\psi\rangle) \\ & \wedge (\langle\psi|\Pi_{(a,j)}\Pi_{(b,k+1)}\Pi_{(a,j)}|\psi\rangle < \langle\psi|\Pi_{(b,k+1)}|\psi\rangle)\end{aligned} \tag{4.79}$$

Vectors for which a primitive inclusion relation holds span a subspace, which can be represented by a projector $\Pi_{(a,j)\succcurlyeq(b,k)}$, defined as follows:

$$\begin{gathered}\left(\Pi_{(a,j)} \succcurlyeq_{|\psi\rangle} \Pi_{(b,k)}\right) \iff \left(\Pi_{(a,j)\succcurlyeq(b,k)}|\Psi\rangle = |\Psi\rangle\right) \\ \Pi_{(a,j)\succcurlyeq(b,k)} = \left(1 - \Pi_{(a,j-1)}\right)\end{gathered} \tag{4.80}$$

4.9.2 Quantum Representation of Documents

If we stick to the analogy with Quantum Theory, documents would be represented by positively defined hermitian operators with unitary trace. The meaning of the density operator can be derived from Gleason's theorem [121], and it is that of a probability measure: that is, a function assigning a number between 0 and 1 to every projector that can be interpreted as the probability of a certain measurement outcome, given a certain preparation procedure (state) of a physical system.

If SEs are represented as projectors $\Pi_{(t,w)}$, a document would be represented by a density operator ρ_D that gives the right probabilities of preservation, however they are defined. One way is, for example, taking the probability of preserving a position of a given document chosen at random with a particular *a priori* bias. If the bias is determined by the identity of the term, there is an *a priori* probability $P_0(t)$ of picking a term t, and this would define a weighted norm like those defined in 4.7:

$$P(t \text{ preserved by } E(t^*, w^*)) = \frac{|E(t^*, w^*) \cdot D|_{\{P_0(t_i)\}}}{|D|_{\{P_0(t_i)\}}} = \frac{\sum_i P(t_i)|E(t_i, 0)E(t^*, w^*) \cdot D|}{\sum_i P(t_i)|E(t_i, 0) \cdot D|} \tag{4.81}$$

$$(\rho_D)_{j,k} = \frac{\sqrt{P_0(t_i)P_0(t_j)}|([(1+i)E(t_j, w_{t_j}) \circ E(t_k, w_{t_k}) + (1-i)E(t_k, w_{t_k}) \circ E(t_j, w_{t_j})] \cdot D|)}{2\sum_t P_0(t)|E(t, w_t) \cdot D|} \tag{4.82}$$

It has to be noted that expression 4.82 cannot be immediately expressed in terms of dependence on latent variables as was discussed in section 2.3.2 (equation (2.11)), since the dependence of two different terms is in the product of SEs; co-occurrence is here included in a more direct way.

4.10 Summary

In this chapter, the notion of Selective Eraser (SE) was defined as a way of considering lexical measurements in a similar way that measurements of physical quantities are performed on physical systems. From the very definition, two ways of applying this concept were presented:

- Definition of a family of norms for documents (section 4.2.2). This rather formal and theoretical concept was revisited in 4.5.1, and 4.8.2 as part of the information given by the term-term kernels that can be obtained from lexical measurements with SEs.
- Extraction of useful features from natural language text that are used in several techniques. In particular in section 4.3, a novel kind of lexical measurements is introduced, based on relations between SEs as they operate on text: *critical widths*. These are explored as a way of representing text documents that go beyond bag-of-words approaches, and is closely related to methodologies based in distances between occurrences.
- Some ways of using SEs to obtain and process information about co-occurrence of different terms is also explored in section 4.5.1.
- The relation of SE-based measurements with probabilistic concepts is explained in section 4.7.
- Mathematical relations between SEs were also explored in different sections of this chapter, which will be used to define transformations in next chapter. In section 4.6 non-Boolean logical relations are defined and explored, and in section 4.8.2 the foundations are laid for a linear algebra of SEs, wich will be the starting point for the definition of further concepts.

Chapter 5

The Aboutness Witness (AW)

In chapter 4 a scheme for lexical measurements was proposed that is inspired in measurement as it is described in physics, and some applications of this scheme for examining and characterising natural language text were outlined. However, up to this point, no concrete proposal has been put forward on how to build a tool for retrieval tasks based on these considerations. In this chapter, we attempt the definition and characterisation of operators that will be sensitive to semantic contents, and can be thus used for retrieval tasks. In section 5.1, the nature of this operator is briefly presented as an analogy to the Entanglement Witness defined an used in QT (a deeper discussion of the quantum case can be found in appendix E). Operators and documents are here represented in a Hilbert space for illustrative reasons, so the analogy to QT can be clear. In section 5.2 we introduce the concept of *Lexical Neighbourhood*, as a way to apply linear combinations of SEs to match particular patterns in the use of terms. In section 5.3 we propose a simple way to build AW from a text query and a text repository. In section 5.4 we examine wether the witness defined in the last section reflects the characteristics of aboutness outlined in section 2.2. Finally, in section 5.5 we summarise the characteristics of the witness defined and outline further developments and applications.

5.1 Discriminating Operators from a Quantum Analogy

After having shaped a measurement scheme after that described by QT and adopted the mathematical tools this theory uses (linear algebra, Hilbert spaces), the idea and name of the witness is comparatively a small loan. The analogy that leads to name it is not as deep and far-reaching as those from measurement, so the focus in this chapter will be on the derivation of the concepts and not in the analogy itself. However, let us start with some words about the idea that motivated the name of the Aboutness Witness.

In Quantum Information Theory there is no such concept as *aboutness*, or anything similar, but we have taken the name "Witness" from an operator that is defined to recognise *entangled* states of a composite system (as opposed to *separable* states) [124] (see an explanation of entanglement and entanglement witnesses in appendix E). We do not claim any further analogies beyond that of a linear operator tailored to perform a classification task, so the nature of the problem of distinguishing entangled states is not relevant for this work.

An operator is defined as a transformation on the elements of a vector space. *Linear* operators, are a particular kind of transformations for which operations such as multiplication by scalars, sum, and therefore linear combination are defined. The use of linear operators follows then naturally from the analogy of considering documents as elements in a vector space, and measurements as transformations acting on them. To introduce the use of a linear operator as a classifying device, it is important to revisit the possibility of defining a linear algebra of SEs introduced in section 4.8.

A linear operator can be said to *scale* an object if it changes its norm. We can define a combination of SEs as acting in a linear way on documents, as follows:

$$|(\alpha E(t_1, w_1) + \beta E(t_2, w_2))D| = \alpha|E(t_1, w_1)D| + \beta|E(t_2, w_2)D| \tag{5.1}$$

Operator-based classification is not a new idea; there is a whole class of linear classifiers that can be expressed as an operator [125] (*Support Vector Machines* are perhaps the most known representative of this class [126]). What is new in this work is that we will not use an explicit vector representation for the objects to be classified (documents) but will compute a score from the

application of a witness.

The basic idea of an AW W_{topic} is that it can be used in a simple way to assess the degree to which a document D is about the topic:

$$\mathrm{A}(D \boxdot\!\!\rightsquigarrow [topic]) \propto |W_{topic}D| \tag{5.2}$$

where $\mathrm{A}(D \boxdot\!\!\rightsquigarrow [topic])$ is the degree of aboutness and $\propto$ means "proportional to".

In the next section, it will be shown that a particular linear combination W of SEs can indeed assess the degree of aboutness of a document to another. Then, in the next section, we will discuss how to build an AW from a few query terms, instead of from a document.

5.2 Lexical Neighbourhood of a Keyword

Let us focus on an *ad hoc* retrieval task where the user provides a few *keywords* chosen to define a particular topic, perhaps in an incomplete and loose way. The classifying operator that will be defined here is designed to perceive not only occurrences of terms, but also tendencies of different terms to occur at certain distances. In section 4.3 we showed that distances between occurrences can be determined by measurements with SEs with different window widths; now we will extend that to define a profile for the neighbourhood of a term where a keyword is likely to be. The precise definition of this profile is:

Definition 5.1 (Lexical Neighbouring Profile)
The Lexical Neighbouring Profile Φ of a term t is a function that assigns a number $\phi(w)$ to each position in a text sequence, according to its distance w to the closest occurrence of t. It will be noted as $\Phi(t)$.

It is worth stressing that the assignation of numbers to places in the sequence does not take the positions as absolute, but relative to the nearest occurrence of the central term. For that reason, it is length-independent, and local in nature, which means that how it works on an interval of the text sequence of the document only depends on the nearby terms. This rules out direct dependences on the length of the document.

In section 4.3.2 a way of finding a distribution of distances was discussed, and this turns out to be a particular case of lexical neighbouring profile. Note that operator $E(t,w) - (1-\delta_{w,0})E(t,w-1)$ (where $\delta_{a,b}$ is the Kronecker delta, that is 1 if $a = b$ and 0 otherwise) preserves precisely the terms that are at a distance of exactly w from the closest occurrence of term t, so we can use it to give $\Phi(t)$ an explicit expression in terms of SEs:

$$\Phi(t) = \sum_{w=0}^{w_{max}} \phi_t(w)(E(t,w)-(1-\delta_{w,0})E(t,w-1)) = \phi(0)E(t,0)+\sum_{w=1}^{w_{max}} \phi_t(w)(E(t,w)-E(t,w-1)) \tag{5.3}$$

The way how a profile is matched by a combination of SEs is shown with an example in figure 5.1.

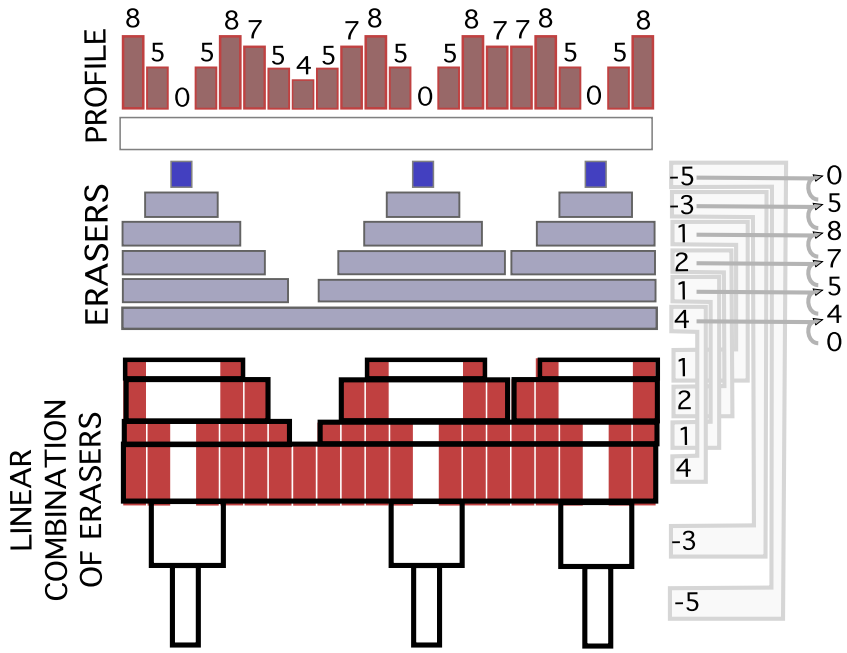

Figure 5.1: Combination of SEs that match a lexical neighbourhood profile. A combination of six SEs with the appropriate weights can reproduce the profile shown in the upper part of the graph. The combination $W = -5E(t,0) - 3E(t,1) + E(t,2) + 2E(t,3) + E(t,4) + 4E(t,5)$ would scale the different positions around t in 0, 5, 8, 7, 5 and 4 according to the distance to the central term.

To build the AW with the profiles, we will use, again, a linear combination:

$$W = \sum_t \alpha_t \phi(t) = \sum_t \alpha_t \phi_t(w) \left(\sum_{w=0}^{w_{max}} \phi_t(w)(E(t,w) - (1-\delta_{w,0})E(t,w-1)) \right) \tag{5.4}$$

where α_t is a weight assigned to term t. Note that this term weight α_t must only be considered separately if the profiles are normalised.

How the profile itself is built, is open to different possibilities. The most obvious way is statistical: the profile can be simply the probability that the keyword actually appears at a given length. It can also be weighted by the probability that a term related to the topic appear at a given distance, with a given probability for some terms to be related to the topic. It can also be weighted by distances, giving a different scaling factor to different distances (for example, to favour smaller distances).

There are, however, other typses of profiles we can use, including term-dependent, position-dependent. To use the full power of a quantum-inspired approach, a scheme can be build based on the use of complex profiles (this is outlined in appendix F).

5.3 Procedure to Obtain an AW for a Query

The simplest witness we may think of is that made with zero-width SEs centred on keywords. This is equivalent to make a TF scoring of documents, and with an appropriate term-weighting scheme like those described in section 4.2.2, it can become TFIDF. We can go one step further, and start including other terms that are associated to the keywords, reproducing their lexical neighbourhood profile, with combinations of SEs like those described in section 5.2. Whatever the evaluation of the norm, the steps to build the witness are:

1. Find the best set of ancillary terms
2. Evaluate the lexical neighbouring profile for each
3. Implement a way of measuring in how much of the document the different profiles and possibly the keyword terms coincide (this can be a nonlinear procedure, but has to do with the norm and not with the application of the witness).

5.3.1 Terms to Build an AW

Given our starting point of a text query submitted by the user, we have a set of keywords (query) to start building the AW. They should describe the topic, but it is possible that a document that treats the topic does not contain any of them. It is convenient to add more terms (which we will call *ancillary terms*) to the description, choosing those who would tend to co-occur with the keywords but are not too common in the collection. The best procedure to choose them is to compute their profiles first as is suggested in the next subsection, and then take the sum of the profiles for all widths as an overall score for each term:

$$\alpha_t = \sum_w \phi_t(w) \tag{5.5}$$

Observe that we have used the same symbol α_t as the weight in the witness; this amounts to claiming that we compute a huge witness with all of the terms co-occurring with the keywords, but then truncate it taking the most important terms.

5.3.2 Lexical Profiles

A simple way of obtaining suitable lexical profiles from a whole collection or corpus is assigning to each keyword a score, and compute the fraction between a weighted token counting for that distance, and a total token counting:

$$\phi_t(w) = \frac{1}{N}\left(\sum_{(|E(t,w)D|-\delta_{w,0}|E(t,w-1)D|)>0} \frac{|E(t,w)D|_{\{q_i\}} - \delta_{w,0}|E(t,w-1)D|_{\{q_i\}}}{|E(t,w)D| - \delta_{w,0}|E(t,w-1)D|}\right) \tag{5.6}$$

where the sum is made over all documents where $(|E(t,w)D| - \delta_{w,0}|E(t,w-1)D|) > 0$, and N is the number of documents that were taken into account. $|\cdot|_{\{q_i\}}$ denotes a weighted norm, like those defined in section 4.2.2; a weight zero is assigned to every non-keyword term, and some sensible weights $\{q_i\}$ are assigned to query terms. These weights are discussed in the next subsection.

5.3.3 Norms and Term Weights for AW

For the values of the profile in (5.6) to be between 0 and 1, we only need to ensure that the query weights $\{q_i\}$ are between 0 qnd 1. We can also fix these weights with simple conditions, such as $\phi_{k_i}(0) = q_i$ for keyword terms. This condition is met when

$$q_i = \frac{\sum_{D_j \text{ with } k_i} |E(k_i, 0)D_j|}{|E(,0)D_j|_{\text{max over all } k} N_{\text{docs with } k_i}} \tag{5.7}$$

With a set of term weights and profiles, a witness can be built easily. Profile coefficients cannot be used directly on SEs, because what the SEs count is cumulative, as can be seen in figure 5.1. For that reason, every SE will be assigned a different factor that is the difference between the weight assigned to this value of w, and the sum that all the profile weights corresponding to larger values of w:

$$W = \sum_t S_t \left(\sum_{w=0}^{w_{max}} \left(\phi_t(w) - \sum_{w'=w+1}^{w_{max}} \phi_t(w') \right) E(t, w) \right) \tag{5.8}$$

5.4 From Uncertain Conditional to Aboutness

Let us first consider the behaviour and relations of SEs alone, and look for concepts that allow us to define uncertain conditional, so we can obtain a way of relating the AW with aboutness, and check if its characteristics comply with what an AW measures. It was shown that a SE can include another; we can jump from there to a definition of uncertain conditional between SEs.

A version of the Ramsey test (section 2.11) can be defined to assess the degree implication between SEs. The Ramsey test consists in measuring how much information has to be added in order to make an implication certain; for SEs it would be as follows:

Definition 5.2 (Uncertain Conditional of SEs within a document)

The degree I of implication between two SE $\text{I}(E(a, w_a) \underset{D}{\rightarrow} P(E(b, w_b)))$ within a document D, where a and b occur is the fraction of the term preserved by a minimal SE $E(c, w_c)$ that includes

both, that is also presrved by $E(a, w_a)$.

$$\mathrm{I}(E(a,w_a) \underset{D}{\rightarrow} E(b,w_b)) = \max_{E(c,w_c)\geqslant E(b,w_b), E(c,w_c)\geqslant E(a,w_a)} \frac{|E(a,w_a)D|}{|E(c,w_c)D|} \tag{5.9}$$

where inclusion relation is defined by $(E_1 \geqslant_D E_2) \iff (|[E_2 \circ E_1]D| = |E_2D|)$ where $\circ$ denotes the composition of SEs. To get rid of the order relation as a constraint, another formula can be used:

$$\mathrm{I}(E(a,w_a) \underset{D}{\rightarrow} E(b,w_b)) = \max_{E(c,w_c)} \frac{|[E(a,w_a) \circ E(c,w_c)]D| \times |[E(b,w_b) \circ E(c,w_c)]D|}{|E(c,w_c)D| \times |E(b,w_b)D|} \tag{5.10}$$

5.4.1 Implication between AWs

For combinations of SEs such as the AW, this relation cannot be used like it is, because reflexive implications such as $W \rightarrow W$ would fail to produce a 1, because they are not necessarily idempotent. However, a simple modification makes it compliant with the reflexive rule:

$$\mathrm{I}(W_a \underset{D}{\rightarrow} W_b) = \max_{W_c} \frac{|[W_c \circ W_a]D| \times |[W_c \circ W_b]D|}{|[W_c \circ W_c]D| \times |[W_b \circ W_b]D|} \tag{5.11}$$

Intuitively, what an implication $W_1 \rightarrow W_2$ means within a document D, is basically that the positions of the document that contribute to $|W_1D|$ will tend to include the positions of the document that contribute to $|W_2D|$. The formula can only refer to AWs that produce nonzero values of $|WD|$; otherwise, the degree of implication $W_1 \underset{D}{\rightarrow} W_2$ should be zero when only $|W_2D| = 0$, one when only $|W_1 = 0|$ and undetermined when both are zero.

5.4.2 Aboutness and Implication between Documents

At this point, these relations define a kind of uncertain conditional for witnesses, but to relate them to *aboutness* it is necessary to be able to establish relations of implication between documents. This arises from the view that an implication can be used to define aboutness (see 2.17): $(A \boxdot\!\!\rightsquigarrow B) \iff (R(A) \rightarrow R(B))$ where $R(\cdot)$ is some kind of representation where implication

can be defined: in this case, it will be an operator ρ_D; a linear combination of SEs. We assume that

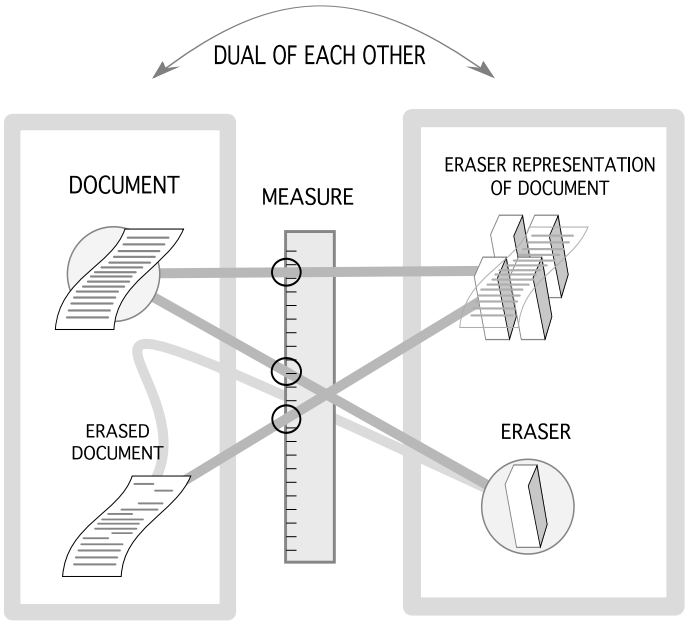

Figure 5.2: Schematic representation of some of the concepts. The measure, a map to a number, mediates between dual objects: things on the side of the erasers are dual to the things on the side of the documents. The document defines a combination of SEs which represent it in the dual space, and SE acts on document to produce a transformed SE which is also in the side of documents.

$$(D_1 \underset{C}{\boxdot\!\!\rightsquigarrow} D_2) \iff \forall D \in C, \rho_{D_1} \underset{D}{\to} \rho_{D_2} \tag{5.12}$$

where C is a collection of documents. Since we are quantifying a degree of implication, we can also quantify the degree of aboutness A, as follows:

$$\mathrm{A}(D_1 \underset{C}{\boxdot\!\!\rightsquigarrow} D_2) = \min_{D \in C} \mathrm{I}(\rho_{D_1} \underset{D}{\to} \rho_{D_2}) \tag{5.13}$$

To translate the definition of implication from the realm of witnesses to that of documents, we can use the concept of *duality*. Duality consists in that documents can be seen as functionals that assign a number to every witness, just as witnesses are considered as functionals that assign a number (score) to every document. A simple scheme of the how documents, erasers, measure and duality are related can be seen in figure 5.2 We can also assign to every document an operator, by a simple formula:

$$\rho_D = \sum_i \chi_i \frac{|D|}{|E_i D|} \times E_i \tag{5.14}$$

where $\{E_i\}$ is a set of SEs. Note that ρ_D is an combination of SEs, and can be considered as an AW: one that detects documents that are about document D. If the coefficients are normalised, then this operator has the nice property that

$$|\rho_D D| = \sum_i \chi_i \frac{|D|}{|E_i D|} |E_i D| = |D| \sum_i \chi_i = |D| \tag{5.15}$$

This may suggest that $\frac{|\rho_D D_x|}{|D_x|}$ is a suitable definition of a degree of aboutness, but we cannot check the properties of this one unless we determine how the coefficients $\{\chi_i\}$ are chosen. Instead, we can check the properties of an aboutness degree by using an analogy of equation (5.11) for documents. We would assume that a third document D_c (which we will call *covering* vector) can be chosen to maximise the value, and evaluate it on an arbitrary document D, as follows:

$$\mathrm{A}(D_a \underset{D}{\boxdot\!\!\rightsquigarrow} D_b) = \max_{D_c} \frac{|[\rho_c \circ \rho_a]D| \times |[\rho_c \circ \rho_b]D|}{|[\rho_c \circ \rho_c]D| \times |[\rho_b \circ \rho_b]D|} \tag{5.16}$$

This can be used precisely as a criterion to choose coefficients $\{\chi_i\}$ by imposing the condition:

$$\frac{|\rho_a D_b|}{|D_b|} = \min_D \left(\max_{D_c} \frac{|[\rho_c \circ \rho_a]D| \times |[\rho_c \circ \rho_b]D|}{|[\rho_c \circ \rho_c]D| \times |[\rho_b \circ \rho_b]D|} \right) \tag{5.17}$$

which is linear in the coefficients $\{\chi_x\}$ for document D_a, but quadratic on those of document D_b and D_c. The evaluation of this expression would, besides, imply a maximisation-minimisation over every pair (D, D_c) of the collection, so its interest is purely theoretical: it shows that with a suitable way of assigning $\{\chi_x\}$ operator $\frac{1}{|D|}\rho_D$ is a valid aboutness witness that can determine in the way suggested in whether a problem document is about D.

1. Reflexivity $A \boxdot\!\!\rightsquigarrow A$ (2.2)

   Every document is about itself. Relation (5.16) gives an implication degree of 1 when all the covering documents documents involved are the same.

2. Transitivity $[A \boxdot\!\!\rightsquigarrow B] \wedge [B \boxdot\!\!\rightsquigarrow C] \Rightarrow [A \boxdot\!\!\rightsquigarrow C]$ (2.3)

   Since the aboutness relation defined here is not binary (about or not about) but fuzzy, a strict

way of interpreting transitivity is by stating that implication degrees are multiplicative:

$$\mathrm{A}(D_1 \underset{D}{\boxdot\!\!\rightsquigarrow} D_2) = \mathrm{A}(D_1 \underset{C}{\boxdot\!\!\rightsquigarrow} D_2)\mathrm{A}(D_2 \underset{C}{\boxdot\!\!\rightsquigarrow} D_3) \tag{5.18}$$

It can be verified that this condition is fulfilled whenever the covering document that maximises I for both implications is the intermediate D_2 (represented by ρ_2):

$$\begin{aligned}\mathrm{A}(D_1 \underset{C}{\boxdot\!\!\rightsquigarrow} D_2)\mathrm{A}(D_2 \underset{C}{\boxdot\!\!\rightsquigarrow} D_3) &= \left(\frac{|[\rho_2 \circ \rho_1]D| \times |[\rho_2 \circ \rho_2]D|}{|[\rho_2 \circ \rho_2]D| \times |[\rho_2 \circ \rho_2]D|}\right)\left(\frac{|[\rho_2 \circ \rho_2]D| \times |[\rho_2 \circ \rho_3]D|}{|[\rho_2 \circ \rho_2]D| \times |[\rho_3 \circ \rho_3]D|}\right) \\ &= \frac{|[\rho_2 \circ \rho_1]D| \times |[\rho_2 \circ \rho_2]D|}{|[\rho_2 \circ \rho_2]D| \times |[\rho_3 \circ \rho_3]D|} = \mathrm{A}(D_1 \underset{C}{\boxdot\!\!\rightsquigarrow} D_3)\end{aligned} \tag{5.19}$$

Implication is also transitive whenever the operators representing the covering vectors for the two implications ρ_c and ρ_d are related through the following expression:

$$\frac{[|\rho_c \circ \rho_2]D| \times |[\rho_d \circ \rho_2]D|}{[|\rho_2 \circ \rho_2]D| \times |[\rho_d \circ \rho_d]D|} = 1 \tag{5.20}$$

This may suggest that the transitive rule does not in general hold for optimal implication degrees, but can hold for suboptimal. For the optimal values of the degree of implication, a weaker relation holds always:

$$\max(\mathrm{A}(D_1 \underset{C}{\boxdot\!\!\rightsquigarrow} D_2), \mathrm{A}(D_2 \underset{C}{\boxdot\!\!\rightsquigarrow} D_3)) \geqslant \mathrm{A}(D_1 \underset{C}{\boxdot\!\!\rightsquigarrow} D_3) \geqslant \mathrm{A}(D_1 \underset{C}{\boxdot\!\!\rightsquigarrow} D_2)\mathrm{A}(D_2 \underset{C}{\boxdot\!\!\rightsquigarrow} D_3) \tag{5.21}$$

3. Set Equivalence (for a given equivalence relation $\equiv$) (2.4)

$$[A \boxdot\!\!\rightsquigarrow B] \wedge [B \equiv C] \Rightarrow [A \boxdot\!\!\rightsquigarrow C] \qquad [A \boxdot\!\!\rightsquigarrow B] \wedge [A \equiv C] \Rightarrow [C \boxdot\!\!\rightsquigarrow B] \tag{5.22}$$

This relation is fulfilled trivially if the following definition is adopted for equivalence:

$$(D_1 \equiv_D D_2) \iff (|[\rho_1 \circ \rho_2]D| = |[\rho_1 \circ \rho_1]D| = |[\rho_2 \circ \rho_2]D|) \tag{5.23}$$

4. Left Monotonic Union (for a given operation of union of sets of infons $\cup$) (2.5) and Cut (also

for a union) (2.6)

$$[A \boxdot\!\!\rightsquigarrow B] \Rightarrow [A \cup C \boxdot\!\!\rightsquigarrow B] \qquad [A \cup B \boxdot\!\!\rightsquigarrow C] \wedge [A \boxdot\!\!\rightsquigarrow B] \Rightarrow [A \boxdot\!\!\rightsquigarrow C] \quad (5.24)$$

These two expressions can be used to actually define the operator representing *the union of documents*

5.5 Summary

In this chapter the Aboutness Witness was defined: an operator that scores a document according to the degree in which it is about a topic. This is developed from the idea of uncertain conditional discussed in 2.4 as an operator that checks if a document is about another document; then, a methodology is proposed to build it from a small set of query terms representing a topic. The properties of aboutness presented in section 2.2 are reviewed to ensure that the AW complies with them, making a case for its theoretical correctness.

An AW should be sensitive to lexical features of the use of terms in documents, like occurrences and distances between them. To achieve this, two aspects are considered in the construction of an AW: term weighting, and a lexical profile. The latter is the most practically innovative part of the methodology: it consists in a weighting of positions around the occurrence of a set of central terms, according to their distance. This is assumed to catch the lexical features that can account at least in part for the degree of aboutness.

This chapter has two most important outcomes: the first is a formal definition of the degree of aboutness between documents within a collection that follows from the idea of measurement, and uses some concepts of linear algebra and the notion of uncertain conditional. This definition comes from considering the linear space of SEs (the space spanned by their linear combinations) as a dual space to that of documents: an appealing idea that will probably trigger more research on both the theoretical and the experimental aspects of its use. The second outcome is a practical, concrete proposal to apply this idea to a retrieval scenario of an *ad hoc* task.

Chapter 6

Ad hoc retrieval with Aboutness Witness

In order to check the applicability of some of the ideas developed in this work, we will perform some standard evaluations of the Aboutness Witness in an *ad hoc* retrieval scenario. 50 assessed topics and five collections of the TREC-1 initiative were used for this evaluation.

6.1 Scenario and Task

The basic scenario consists in a constant collection that is queried with a series of *ad hoc* queries submitted by the user in the form of a set of a few terms [26, page 137]. From these queries, an AW is built and used to score the documents and generate a ranking.

The AW is built by completing the set of query terms with a number of ancillary terms chosen from the collection, assigning the SEs centred on all the terms coefficients that will be determined from with co-occurrence information. This information is gathered from the from the whole collection without any ranking of documents involved.

Since terms with a non-topical role in language (stopwords) account for most of the occurrences in a given document, we removed them at indexing time, so the sequences of text examined at retrieval time did not contain stopwords. This is also likely to shorten the distances between occurrences, letting more of them fall within a limited set of distances below a fixed maximum width.

For each collection, a list of stopwords was made based on number of occurrences in the whole

collection $N_{t\ in\ C}$ (should be high for a stopword) and average covering width $W_{t\ inD}$ (see definition 4.9) as a score $N_{t\ in\ C}/Wt\ \bar{in}\ D$, and then selected by hand. This is based on an assumption on distribution of occurrences in text made in section 4.3.2. Keywords in queries were also removed.

6.1.1 Methodology and Evaluation

Two main parameters were fixed for each run of the method: number of terms, and window width. All queries had to be augmented in order to take advantage of co-occurrence information, up to 10 and 15 terms (fixing a number that is smaller than the length of some of the queries may cause trouble). A maximum width was fixed for the witness, so that the profiles were considered from width zero to that maximum.

The steps followed by the program to produce the ranking of documents for a topic were the following:

1. Assign preliminary weights to query terms. This would help favouring terms that co-occur with rare query terms, to compensate for the potential absence of these rare terms in relevant documents.

2. Gather a large set of term lexical profiles, as weighted distributions of probabilities that a query term appears at a given distance for the considered term. Profiles are obtained for all terms co-occurring with the query terms within the distances considered. A weight is assigned to every distance between nearby occurrences from distance 0 to a maximum distance w_{max}. Weighted Co-occurrence counts are used both for building a lexical profile and for adding up a score for the term.

3. Choose the higher scoring terms, together with their profiles. In the set of chosen Terms, scores are normalised over the whole set such that the highest is one, and profiles are normalised for each term to sum up to one.

4. Compute the score by applying the witness to each document. The witness will assign a number to each position, cumulating score from all the lexical profiles. Then, a power of this cumulated score was summed to the overall score. The exponent of this power was chosen to give more weight to overlap.

For sets of terms with similar scores, the highest scored positions will be those falling in the overlap of several SEs from the AW, whil for sets of terms with very dissimilar scores, the highest scores will tend to simply reflect the presence of a highly scored term. In the first case, using alarge exponent for the cumulated score will give more importance to overlap, while in the second case, a high power would simply increase the importance of the presence of an already highly scored term. For this reason, the chosen exponent was the sum of scores: since the highest is fixed to 1, this exponent will be high for sets of terms with similar scores, and low when the scores are dissimilar.

The contribution of each position in the document would come from formula (5.8) and would be:

$$|WD|(\text{position } i) = \sum_t S_t \left(\sum_{w=0}^{w_{max}} \left(\phi_t(w) - \sum_{w'=w+1}^{w_{max}} \phi_t(w') \right) |E(t,w)D|_{\text{position } i} \right) \tag{6.1}$$

where S_t are individual term weights, and $\phi_t(w)$ are the lexical profiles. The overall score is:

$$|WD| = \sum_{\text{position } i} (|WD|(\text{position } i))^{(\sum_t S_t)} \tag{6.2}$$

Two main performance measures were used to assess the performance of the method: Mean Average Precision and BPref, described in the next subsection. **Mean Average Precision** (MAP) is computed by scanning the ranking of documents top down, and computing the precision value for the interval of documents from the top 1 to each occurrence of a relevant document. For a given set of rankings for relevant documents $\{R_i\}$ the expression for MAP will be:

$$MAP(\{R_i\}) = \sum_{i=1}^{N_R} \frac{i}{R_i N_R} \tag{6.3}$$

where N_R is the total number of relevant documents. MAP is the most used performance measure. MAP will show very low values for topics with small numbers of assessed documents, because it assumes that every non-assessed document is non-relevant. For that reason, in those cases it is useful to use other measure. We will use BPref, which is similar to MAP but counts only assessed terms. For a ranking of documents with the relevant documents in positions $\{AR_i\}$ and assessed

collection	AP89	WSJ8789
No. documents (x 1000)	84.68	98.73
No. terms (x 1000)	207.62	169.34
No. tokens (x 1 000 000)	41.80	43.68
avg. length	493.66	442.44
avg. covering width	432.27	661.75
avg. covering %	72.71	70.33
avg. docs with term	104.289	127.79
avg. occurrences	1.38	1.46
After stopword removal		
Stopwords used	123	94
No. terms (x 1000)	207.49	169.24
No. tokens (x 1 000 000)	24.25	26.85
avg. length	286.45	271.99
avg. covering width	81.91	534.11
avg. covering %	75.39	91.65
avg. docs with term	81.31	105.30
avg. occurrences	1.28	1.46

Table 6.1: Collections Used for the experiments. Covering width is the minimum width of a SE centred in the term that preserve a whole document, and percentage of covering (covering %) is the covering width as a percentage of the length of the document. T

documents in positions $\{A_i\}$, the formula is:

$$BPref(\{A_i\}, \{AR_i\}) = \sum_{i=1}^{N_R} \frac{iAR_i}{\min(R_i, AR_i - R_i)N_R} \tag{6.4}$$

where R_i is the number of relevant documents in ranks higher or equal to i, and $\{AR_i\}$ is a set of binary numbers, each corresponding to and assessed document i: 1 when it is relevant, and 0 when it is non-relevant.

Collections used

Two collections were used to asses the method, both from the TREC-1 dataset gathered and assessed by NIST [127].

1. **Informal Notes on the Associated Press Newswire, 1989 (AP89)** The material includes copyrighted stories from the AP Newswire, as collected by AT & T Bell Laboratories. The

stories are all from 1989.

2. **Informal Notes on the Wall Street Journal 1987 to 1989 (WSJ 87-89)**

   The material includes copyrighted stories from the Wall Street Journal, mostly from years 1987 and 1988, but also some from 1989 and has been provided courtesy of Dow Jones Information Services.

6.2 results

The method for *ad hoc* retrieval based on AW outlined in chapter 5 and described in a more concrete way in section 6.1 had, in general, a performance that falls in the same interval of measures defined by the four considered baseline methods (TF, TFIDF, BM25 and Language Model). A summary of the results can be found in table 6.2

Collection	AP89			WSJ8789		
Widths	4	10	avg.	4	10	avg.
bpref for AW (10 terms)	26.88%	27.19%	26.79%	25.48%	24.97	25.06%
bpref for AW (15 terms)	26.86%	27.40%	26.80%	24.20%	23.60%	23.83%
bpref with TF	23.81%			20.42%		
bpref with TFIDF	29.96%			29.98%		
bpref with BM25	29.96%			29.98%		
bpref with LMDP	28.86%			24.09%		
MAP for AW (10 terms)	13.46%	13.65%	14.15%	13.67%	13.54%	12.80%
MAP for AW (15 terms)	13.66%	13.52%	13.75%	14.15%	13.85%	13.04%
MAP for TF	9.88%			7.16%		
MAP with TFIDF	14.54%			13.99%		
MAP for BM25	18.94%			16.79%		
MAP for LMDP	14.33%			13.09%		

Table 6.2: General results for the evaluation of the AW as a method for *ad hoc* retrieval

6.2.1 Comparison between the AW and baseline methods topic by topic

To the initial terms provided by the queries, a new set was added to complete 10 terms. The AW outperformed all the baseline methods in most of the topics, as can be seen for bpref in figure 6.1

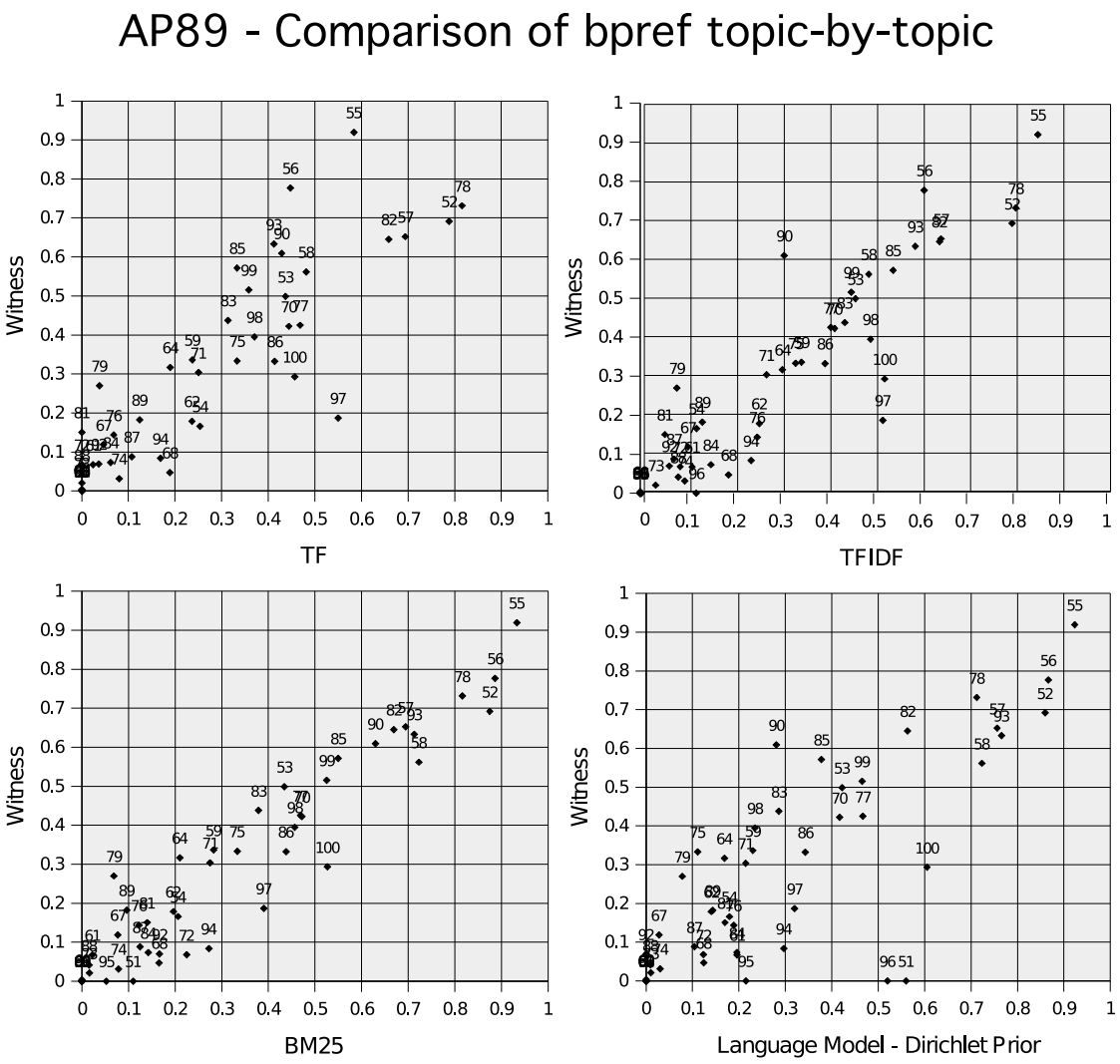

Figure 6.1: Comparison of bpref values obtained with the AW with baseline methods topic-by-topic, for collection AP89. Axis x corresponds to the bprefs obtained by the baseline method, and axis y to that obtained with the AW

for collection AP88, and in figure 6.2 for collection WSJ8789. For collection AP89 a remarkable correlation can be seen with method BM25. The reason for this could be the similarity of the BM25 ranking function and that of a wide-window SE, discussed in section 4.3.1. The fact that the correlation is less strong for collection WSJ8789 supports that statement, since covering width is much larger in the latter (534.11 for WSJ8789 vs 81.91 for AP89 in average). This means that overlap is probably more prevalent (or is prevalent at smaller distances) for AP89, making the resemblance between BM25 and a wide-width SE measurement closer.

Variation of the maximum width for the witness does not introduce an important change in perfor-

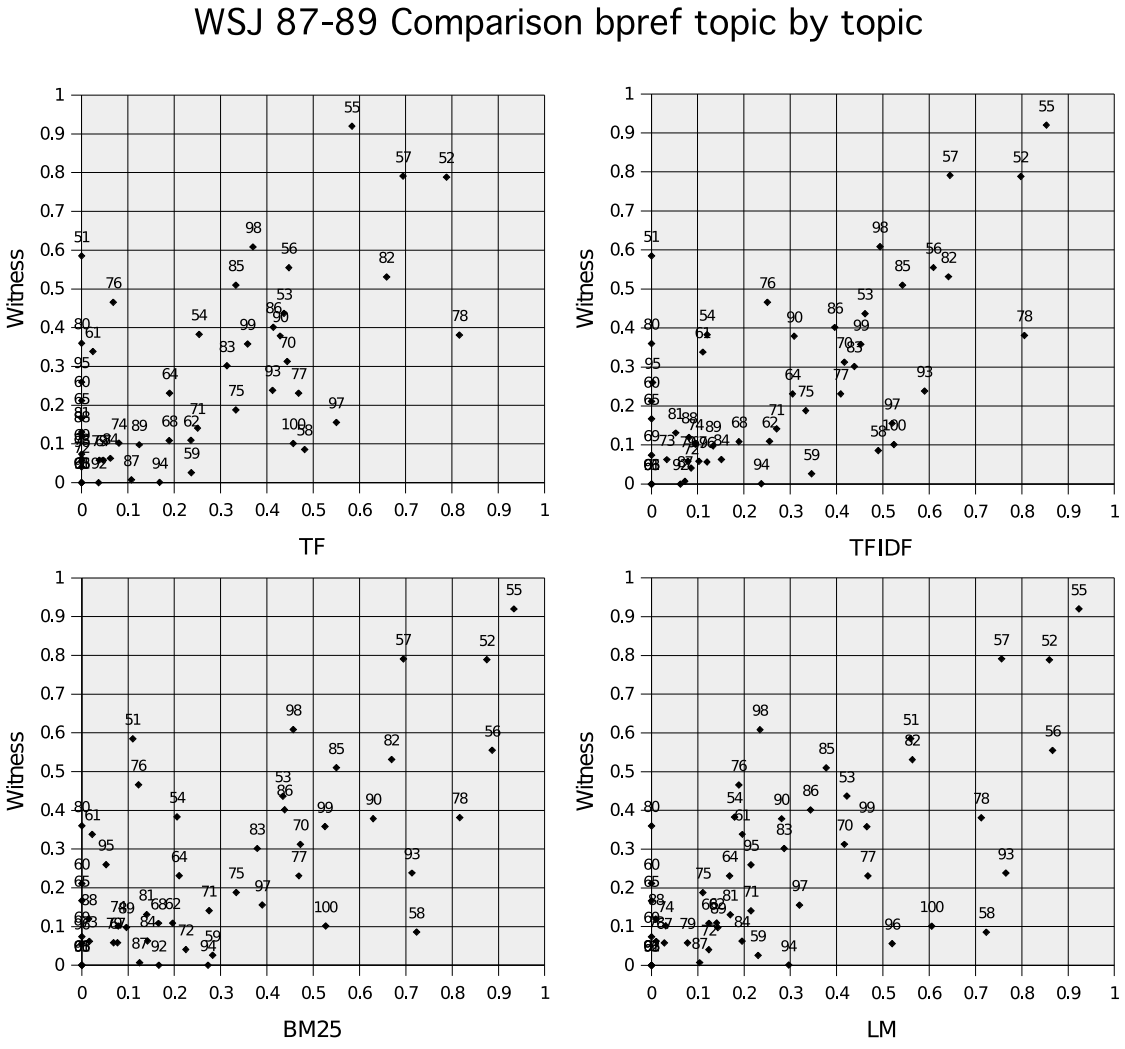

Figure 6.2: Comparison of bpref values obtained with the AW with baseline methods topic-by-topic, for collection WSJ8789. Axis x corresponds to the bprefs obtained by the baseline method, and axis y to that obtained with the AW

mance. In figure 6.3 the change of performance both in bpref and in MAP is shown for different values of maximum width. For collection AP89 there seem to be an optimal range of widths between 2 and 5, while for WSJ8789 performance increases in a more or less monotonic way while increasing maximum width.

6.2.2 Effect of the number of terms used

Since we are using a fixed number of terms to build the AW, it is necessary to check how the number used affects the results. The comparison between the results for 10 and 15 terms, suggests

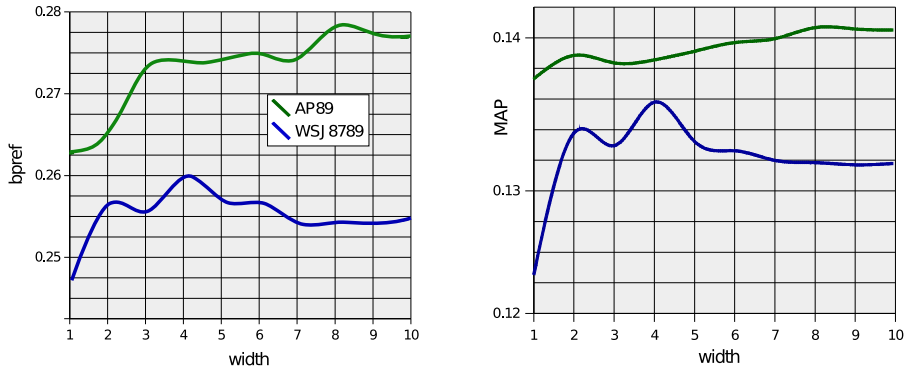

Figure 6.3: Variation of bpref and MAP (averaged on all topics and on all number of terms) with window width for both collections. Note that the scale is quite stretched in axis y to make the variation visible.

that performance is quite robust with respect to this parameter. A linear regression run on all bpref values for topic and collection defining variable x as the value for 10 terms and y the value for 15 terms, show an almost perfect correlation both for bpref and MAP. The results can be seen in table 6.2.2, and the points are plotted in figure 6.4.

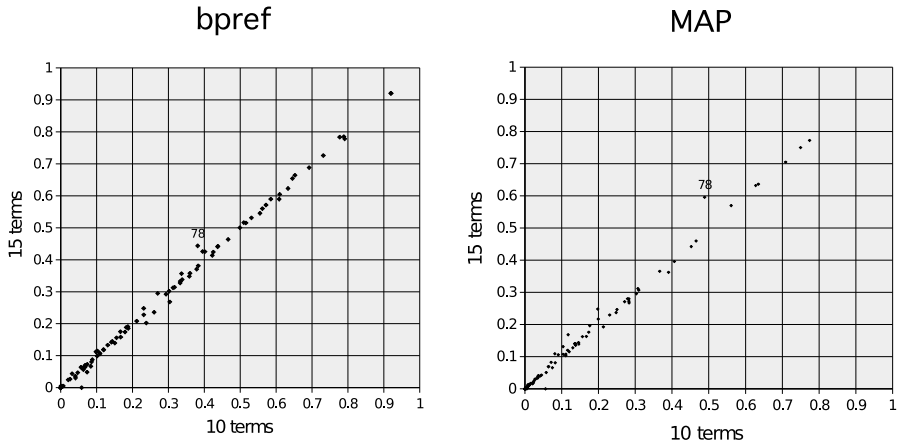

Figure 6.4: Performance with 10 terms vs performance with 15 terms. The outlier point is topic 78 as evaluated in collection WSJ 87-89.

Topic 78 when evaluated in collection WSJ 87-89 seems to have a peculiar behaviour, since augmenting maximum width does increase the retrieval performance notably. The query for the topic is "**Greenpeace**". The Aboutness Witness obtained for it with maximum width 5 is decribed in table 6.2.2 for 15 and for 20 terms. The difference between the AW for 10 and 15 terms is not dramatic, and amounts to adding terms with a very simple profile: one counting only distance 1

Measure	bpref	MAP
Correlation (r^2)	99.71%	99.27%
slope	0.9943	1.0071
cut with 0	0.0020	0.0003

Table 6.3: Results of a linear regression between topic-collection performance values for 10 terms and 15 terms.

from the term. There are only 4 documents marked as relevant for this topic: WSJ861203-0100, WSJ870911-0086, WSJ870924-0017 and WSJ870928-0092. The text surrounding the new terms (those added by the 15-term setting as compared with 10-term) could tell us something about the improvement of performance (hyphens separate different chunks taken from the documents):

For "**stopped**:"

> *has been limited or stopped greenpeace the environmental group*

For "**officials**"

> *of dioxin discharges epa officials said they are moving – that direction currently epa officials said a dow chemical – year convicted two senior officials of one of the – federal investigation waste management officials also called the los – determine whether corporate level officials at waste management and – waste management or its officials going all the way*

Term "stopped" co-occurred very closed in the text with the query term ("greenpeace") and with other terms used in the AW; this is something the method scores quite high. Term "officials" does not co-occur with important terms, but all the terms that surround its occurrences seem to be quite particular of this topic; they might as well be used as central terms.

6.2.3 Topics where the method outperformed baselines

In figure 6.5 the difference in average performances both in bpref and MAP are shown for all the topics, ordered by the difference itself. It is clear in the graphics that the method outperformed

Term	α	$\phi(0)$	$\phi(1)$	$\phi(2)$	$\phi(3)$	$\phi(4)$
with 10 terms						
greenpeace	1.00	1	0	0	0	0
group	0.50	0	0.65	0.20	0.15	0
environmental	0.43	0	0.23	0.67	0.10	0
usa	0.30	0	1	0	0	0
ship	0.19	0	0.45	0	0.23	0.33
council	0.17	0	0.81	0	0.19	0
past	0.14	0	0.75	0	0	0.25
protest	0.14	0	0	0	1	0
flustered	0.13	0	1	0	0	0
with 15 terms						
greenpeace	1.00	1	0	0	0	0
group	0.50	0	0.65	0.20	0.15	0
environmental	0.43	0	0.23	0.67	0.10	0
usa	0.30	0	1	0	0	0
ship	0.19	0	0.45	0	0.23	0.33
council	0.17	0	0.81	0	0.19	0
past	0.14	0	0.75	0	0	0.25
protest	0.14	0	0	0	1	0
stopped (*)	0.13	0	1	0	0	0
flustered	0.13	0	1	0	0	0
sane (*)	0.13	0	1	0	0	0
protested (*)	0.13	0	1	0	0	0
washington (*)	0. 13	0	1	0	0	0
officials (*)	0.13	0	1	0	0	0

Table 6.4: Witness for topic 78 (greenpeace) in collection WSJ 87-89 with maximum width 4. The terms marked with (*) are not in the AW made with 10 terms.

the baselines in roughly half of the topics. The average baseline was outperformed for 3 topics with more than 15% of difference in bpref, and the average baseline outperformed our method in 5 topics by more than 15%. In terms of MAP, the average baseline was outperformed by more than 15% also in the same topics, and outperformed our method by more than 15% in only two topics.

Topics for Best Comparative Performance

Four of the topics where the AW outperformed the average baseline by more than 20% were:

1. **Topic 79** (difference of 29.12% in bpref, 2.52% in MAP) “frg political party positions”

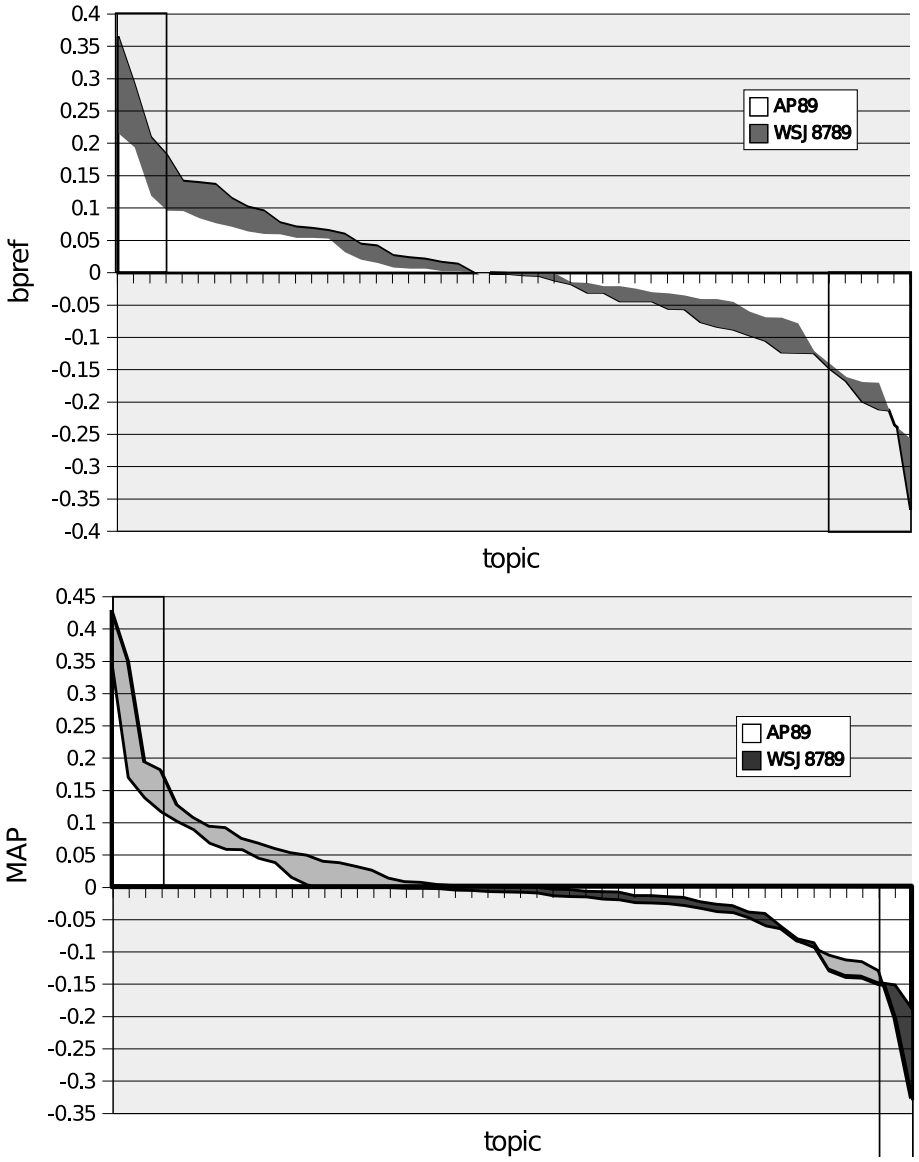

Figure 6.5: Difference of performance between AW (averaged over maximum widths and number of terms) and the average of the baselines (TF, IDF, BM25 and Language Model with Dirichlet prior). The boxes at the sides show with how many of the topics the differences are more than 15%.

"FRG" appears only twice in AP89 and none in WSJ 87-89, while the other keywords are very frequent: the least frequent from them is "positions", which occurs in 1903 documents in AP89 and in 3050 documents in WSJ 87-89. "Party" tends to occur an average of 3.14 times in documents that contain it in AP89, and 2.4870 times in WSJ 87-89, so it will provide rich co-occurrence information.

2. **Topic 64** (difference of 13.99% in bpref, 16.51% in MAP) "hostage taking"

This query is composed of a quite rare term "hostage" (1111 docs in AP89, 339 in WSJ 87-89). The term "hostage", however, occurs an average of 2.01 times per document in AP89, and 1.49 times in WSJ 87-89. allowing to gather co-occurrence with other terms to compensate its scarcity.

3. **Topic 90** (difference of 24.39% in bpref, -4.14% in MAP) "data proven reserves oil natural gas"

   For this topic, there are 122 documents The keywords describing this topic are also terms that tend to occur several times in the documents that contain them (except for "proven"). Oil tends to occur 3.68 times in AP89 and 3.14 times in WSJ 87-89. "Gas" and "Reserves" can be expected to be associated also with the former, thus co-occurring an important number of times.

4. **Topic 85** (difference of 16.53% in bpref, 25.93% in MAP) "official corruption"

   This case is slightly more difficult to explain than the others, since the terms occur an average number of times of respectively 1.60 and 1.59 in AP89, and 1.53 and 1.39 in WSJ 87-89. Both are below the average for query terms, which is 1.69. An explanation of the good performance of a method based on co-occurrence could be the ambiguous character of the term "official", which can be used with different meanings. Information from the surroundings of the term will probably disambiguate it better than counting of terms in the whole document.

6.2.4 Some characteristics of the obtained AWs

The scoring procedure for this method, including short-range co-occurrences, is certainly worth examining, but since this is a prototype version, its more non-standard features are more important and interesting. For that reason, in the next subsection only the most novel aspect of the method will be examined: the lexical profiles. The way lexical profiles were generated is quite simple, and can surely be improved greatly, but the obtained profiles already show interesting features.

Lexical Profiles

Lexical profiles are defined in such a way that they score each distance from an ancillary term (from distance 0 to a maximum) according to three factors, each depending on:

1. The identity of the terms found at such distance. Query terms were already assigned a score that was counted every time the query term appeared at the considered distance
2. The distribution of the term found at such distance within the document. Terms in a cluster where their density was higher counted with more weight than occurrence of the term that were evenly distributed through the document.
3. A decaying factor that prevents longer distances to have too large a share in the profile (the further from the occurrence, the more noise can be expected) [110]

The profile averaged on all ancillary terms for collection AP89 and WSJ8789 can be seen in figure 6.6. The most interesting feature found in the profiles, is that in spite of a damping factor that

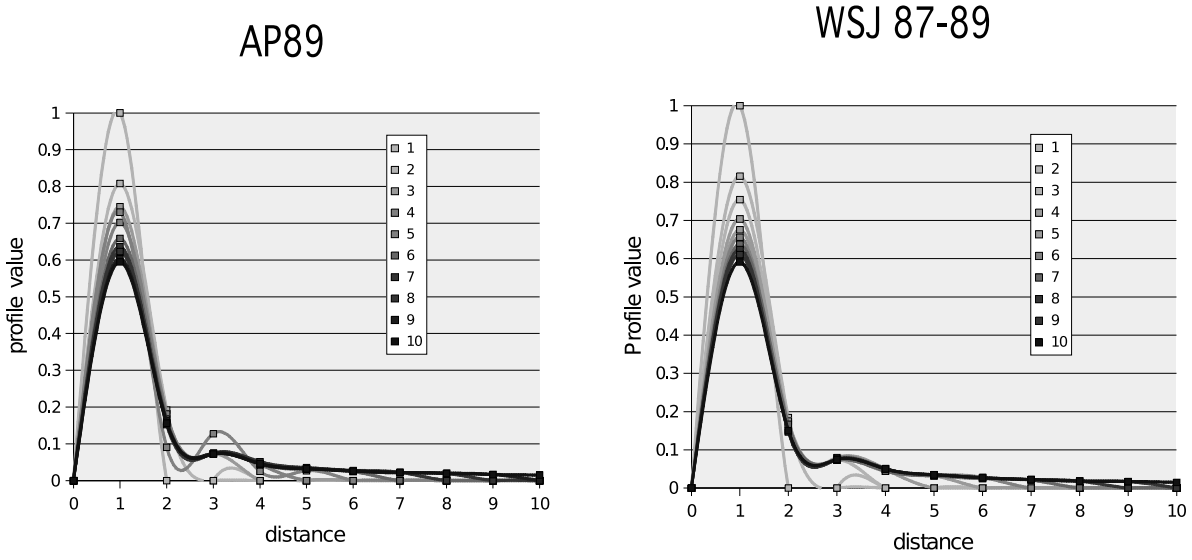

Figure 6.6: Lexical Profiles for ancillary terms as assigned from collections AP89 and WSJ 87-89 with topics 51-100 from TREC-1.

forces the values to go down as distance increases, there is a consistent tendency to form a peak around distance three. This is consistent with earlier studies in co-occurrence, where it has been found that the most semantically significant co-occurrences are within windows of five to eight

terms [50]. The fact that the two profiles for different collections are so similar and that their features are so clear suggest that there is a natural shape for the profile, coming probably from the structure of language itself.

This suggest that the painstaking process of gathering co-occurrence information can be both simplified and perfected by using a universal prior profile, to be modified by the experimental data.

6.3 Summary

Preliminary tests were ran for a methodology based on the Aboutness Witness, an operator defined as a linear combination of SEs that directly assigns a score to every document. There was no substantial improvement in performance with respect to baseline methodologies, but the method proved to be robust with respect to parameter change, as well as sensitive to features of text that are not caught by bag-of-words approaches.

Among the methods used as baselines, the one that behaved in a more similar way to AW was BM25. This joins the evidence gathered in section 4.3 about a close relation between semi-subsumption of occurrence events and measures with wide-width SEs. All the scoring procedures used by the other methods can be adapted to the AW, with some care. A TF version of the method would not use any global weighting of terms, just features from the text sequence and the counts of preserved tokens. TFIDF could assign ancillary terms a factor depending on the number of documents where they co-occur with one of the query terms. BM25 has a scoring function that is intrinsically separated term-by-term, making it difficult to use it with co-occurrences. However, a semi-subsumption analysis ([111]) could mark the way to a generalisation of BM25 to include co-occurrence. For Language Models, the link to AW could come from considering SEs as a very economical representation of n-gram distributions, and treat them as such. The connection to BM25 and Language Models is a whole work in itself, and remains as one of the interesting possibilities that were left aside as a methodological choice.

6.3.1 Strengths of the Method

The AW is not only sensitive to the presence of terms or their counting, but also to the sequence of the text, and the distances between neighbouring occurrences of terms. While co-occurrence in the whole document can be a source of noise, and sliding windows gathering co-occurrences can be quite costly to use, methodologies based on SE are less prone to noise and faster to use at the same time.

Sensitivity to the term sequence in the text can be an advantage for certain tasks where access to a whole collection is restricted leaving the system with very little information to work with, or when the pieces of information to be retrieved are not neatly separated as documents (for example, in passage retrieval).

There are also situations when the problem is not the scarcity of information, but its excess; the method based in AW works with additional criteria to evaluate whether terms are noise or not, and these do not depend on an exhaustive information of the whole collection, or from a large number of documents. This is the case for tasks where entire documents are used as samples along with the queries. In these cases, the AW has an additional advantage: the procedure for obtaining an AW from a document is clearly derivable from the theory, and, as was shown in chapter 5, a document-based AW complies with the conditions of an appropriate quantifier of the degree of aboutness.

6.3.2 Future directions for development of the method

Being a radically new approach, this method can give rise to different new paths for research. Let us mention three of them:

Use of Known Scoring Schemes to Build the AW

It is possible to implement weighting schemes that are known to behave well for retrieval tasks. The methods used here as baselines provide four examples. All these are term-by-term scoring

functions: they split the total score into contribution of the single query terms. An adaptation of the methodology proposed here would be:

$$W = \sum_t W_t = \sum_t \left(\sum_{w=0}^{w_{max}} \alpha_{t,w} E(t,w) \right) \tag{6.5}$$

Four ways of combining the scores produced by the different groups of SEs were implemented, mimicking simple IR methodologies:

1. Term-Frequency:
$$Score = \sum_t |W_t D| \tag{6.6}$$

2. Term-Frequency, Inverse-Document-Frequency (TFIDF)
$$Score = \sum_t X_t |W_t D| \quad \text{where } X_t = \log\left(\frac{N_d - N_{d,t} + 0.5}{N_{d,t} + 0.5}\right) \tag{6.7}$$
where N_d is the total number of documents in the collection, and $N_{d,t}$ is the number of documents containing the term t.

3. Okapi BM25
$$Score = \sum_t X_t \frac{|W_t D|(K_1+1)}{|W_t D| K_1 + b + (1-b)\frac{L_d}{L_{avg}}} \quad \text{where } X_t = \log\left(\frac{N_d - N_{d,t} + 0.5}{N_{d,t} + 0.5}\right) \tag{6.8}$$
L_d is the length of the document, L_{avg} is the average document length in the collection, and parameters K_1 and b are free. In this work they were used with the default values $K_1 = 2.0,\ b = 0.75$.

4. Unigram Language Model with Dirichlet Prior
$$Score = \sum_t \frac{\log(|W_t D| + \mu \quad \underset{\text{averaged in collection}}{|W_t D_i|} \quad)}{\log(|D| + \mu)} \tag{6.9}$$

A problem that arises with these schemes is that there is no way of enhancing the contribution of overlap: there is no way of enhancing it position-by-position because the score is not partitioned

in this way, but term-by-term. This is a problem that remains unsolved, and will be the subject of future research.

Acquisition of Lexical Profiles

The sparsity of co-occurrence data in any collection is a serious problem for most methodologies trying to use it for retrieval, and it is an issue as well for the construction of effective AWs. However, as it can be seen in figure 6.6, it seems that the shape of the profile for topical terms can be quite universal, and it would be possible to use a reasonable prior to enhance enormously our method by a more sophisticated scheme for inference of the parameters.

The study of the profiles themselves seem an interesting subject of research by itself, and there are already a number of studies on the distribution of distances. However, it would be quite useful to perform such studies only on limited sets of terms, like those semantically related to a topic.

Chapter 7

Conclusions

This chapter is a guide to the milestones of the work; a review of the thoughts to take home as a general impression of the thesis. It will be divided in three sections: a short set of remarks on the nature of the work, where its theoretical and fundamental character is briefly discussed. Subsequently, a section where the research questions are revisited, and their answers summarised. Finally, in the third section we describe what is the current status of this work by the end of the thesis, and how it can be expected to go on.

7.1 Remarks about the nature of this work

As it is the case of most Ph. D. projects, in this work several measures were taken to limit the scope of it subject; but thanks to the fundamental character of its basic concepts its nature remained quite general. Some of the conclusions are then very general and even fundamental, but there is also a variety of particular subjects where the development of the overall theme brought quite concrete insights.

The starting point of the work was quite a simple idea: that perhaps the quantum account of measurement and observation is not so strange to humans after all, and can actually be a good way of describing our way of acquiring and processing information in everyday matters. Over time the work took a more concrete and complex shape, and ended up as a fairly large conglomerate of definitions and theoretical results.

As a novel point of view sitting at the basic level of measurement, the proposed approach aims to shed light on operational problems of text retrieval as well as on their theoretical formulation. The work was aimed at this theoretical level, and it is there where most of its results lie. Even though the work is in a sense an *ab initio* (from first principles) endeavour, it provided bridges to connect different kinds of existing models: an example of this are binary relations resembling boolean union, intersection and complement being introduced in Vector Space Models (up to now, quite logics-free) through the lattice of subspaces,

At the same time that abstract motivations guided most of the research, a concrete intent of including co-occurrence of terms into IR in a computationally affordable way was an important driving force, especially for the practical implementations of the work.

7.2 Research Question and their Answers

To sum up the conclusions of this work, let us review the research questions presented in section 1.3. Simple answers were found for most of them, although the mathematics behind of the these answers is less simple.

7.2.1 The Nature of the Lexical Measurement

RQ1 How can basic lexical measurements on documents be defined to match in a very general way the properties of a quantum measurement?

> *As operators acting on documents*. The solution obtained to this question is the concept of *SE*, an operator that acts on text focusing on a particular part of it (tokens surrounding the occurrences of a central term) and erases the information from the rest. Selective Erasers are defined in 4.1.

As an additional asset, SEs are very well suited to relate a number of functions of the occurrence frequencies and distances between occurrences. This is shown in section 4.3 to be useful to extract

a wealth of lexical features from a document, most of them lying beyond the reach of bag-of-words approaches.

Their quantum-inspired origin leaves SEs the definition of several operations involving them, like composition, sum, multiplication by numbers. These linear-algebraic concepts provide the tools to build complex entities with them; and this entities will have perfectly well defined properties and relations (see section 4.8). This fact is used to build operators that can capture lexical features in documents, and represent them numerically.

SEs are related to probabilities through two distinct directions: one operational, through considerations of what their action on documents would produce in different cases (quantification of information erased in section 4.1). Another way to probabilities is more abstract: trhough probabilistic spaces and the concept of *measure* (see section 4.7).

7.2.2 Measuring

RQ2 How can basic lexical measurements capture the features of text that convey meaning?

> *By the use of a complete and coherent mathematical framework to combine basic lexical measurements*. In this work, a path is shown leading from a very basic description of the process of measurement (chapter 3) to an IR methodology. Measurement dictates the properties of SEs (defined and explored in chapter 4), and these properties (together with some inspiration and wisdom from mathematics, physics and, of course, IR) determine how these concepts can be used to build more complex ones that are suitable for use in IR. The fundamental character of the chosen point of view provides the possibility of generate completely new methods (like, for example, the representation of documents with linear combinations of SE in section 5.4) or put together disparate theories an models in a common general framework.

A wealth of more or less complex lexical measurements that can be performed on documents is shown in section 4.3, giving quite interesting insights about the use of terms win written documents, and establishing links to existing methodologies that use lexical measurements for IR.

7.2.3 Linear Operators for IR

RQ3 How to use this approach as a starting point to design better performing IR systems?

> *By summing and multiplying*. SEs constitute a large set of quite variate building blocks, and linear algebra (see section 4.8) provides a set of rules that are flexible enough to produce an enormous amount of different composite operations, while being also stringent enough to keep clear relations and expression for those operations. How these rules and procedures can be used to produce a concept with an immediate application to IR is explained in chapter 5, and can be seen working in chapter 6.

Linear algebra also provides powerful representation techniques, which have been used extensively in IR in a rather heuristic way. Existing schemes such as Vector Space Models are linked to SEs through linear algebra (see section 4.5.2) This link also unifies a number of different approaches, now interpretable as versions of a Selective Eraser approach (for example, co-occurrence based kernels), and suggest new forms of such models. Theere are two fronts where the erasers approach can contribute to IR: representation of documents, and retrieval.

Representation of Documents

In this work, four different and new ways of representing documents have been proposed:

1. By specifying the lattice of SEs: a document would be represented by a set of SEs, plus their order relations (see section 4.4). A particularly simple representation based on that is with a tree of classification of terms based on equivalence classes of increasingly wide SEs (see section 4.4.5. This idea was not developed much further, but bears a relation with uncertain conditional between SEs, and could be revisited with that view.

2. Using a vector space with a measure defined by a kernel built on SEs (see 4.5.2). This representation is close related, but distictly different, than other proposals to use operators acting on Hilbert spaces to represent documents and queries [53].

3. Represenging Documents as Combinations of SEs. This representation was developed in chapter 5 as a way to introduce a formal aboutness relation between documents. This is perhaps the most radically new concept developed in this thesis, and can lead to a wide research field within Information Retrieval but also in related areas like Natural Language Processing.

Retrieval

The concrete method developed for retrieval in this work is the Aboutness Witness, subject of chapter 5. It is a quite general and flexible technique, which can be implemented in a myriad of different ways. In this thesis one implementation was developed and tested. The preliminary evaluation of a working IR methodology in chapter 6 shows that the concepts developed can in fact be applied, and even though the results obtained were not something exceptional, give some additional insight on the way the proposed concepts work in the "real world" of retrieval.

7.2.4 An Encompassing and Unifying Approach

RQ4 Does the point of view proposed (processing of lexical information as a physical measurement) include existing accounts?

> *Indeed*. All term frequency counts can be seen as a special case of the application of SEs plus unerased token count (see 4.3.1). Vector space methods are also quite close to the proposed approach, which can include most of them but also goes far beyond, in terms of both deriving them from basic principles, and including concepts that are absent in them (see 4.5.2). The relation to probabilistic models was discussed in the answer to research question RQ1, and the relation with logical models, perhaps the most fruitful of them all, is threefold: on one hand, it relates to measure and valuation (concepts introduced in section 3.3.1, but used all along the work, especially in 4.7 and 4.2); it also includes extensions of Boolean operations such as "and", "or" and "not", and, finally, provides a starting point for exploring inference-like schemes based on these.

7.2.5 Departing From Classical Logics

RQ5 Does the point of view proposed go beyond existing accounts in a fundamental way?

> Some of the foundations of this and other approaches to IR are still under research and examination, but from the evidence that was gathered throughout this work, we can answer with a ***yes***. Even though we know of proposals to abandon Boolean logic and other well-known defoult frameworks of normal science from more than two decades ago, the research programs that emerged from those calls are few and, in most cases, aim at different targets than this one. Besides, some revolutionary ideas like that of the uncertain conditional were infused new life in the scheme developed.

Aspects of this work that can be considered innovative have been mentioned on the answer to other research questions, but the most fundamental level in which something new can be found, is probably that of three concepts:

- measure, and its relation to duality: how an object of one class can assign an object of another class to a number.
- uncertain conditional, considered as a relation between operators, and accompanied by a functional that assigns a number to this relation: the degree of implication.
- the treatment of transformations on text as operators, and the assignation of mathematical operations and structures to them, like algebras, partial order, and, once again, measures. This allows to combine the former two into a concept that can be applied to IR: the Aboutness Witness.

7.3 The Way Ahead

Being this a very fundamental approach, the new directions it suggests for research are quite wide. A complete scheme of lexical measurement was put together and explored, but there are still a number of theoretical and experimental aspects of the scheme that are still to be studied.

7.3.1 Directions for Theoretical Research

Semi-Subsumption: In section 4.3 an interpretation of the BM25 scoring formula was suggested which used the concept of semi-subsumption [111]: some term occurrences can be considered as being events that are semi-subsumed in other occurrence events. This turned out to have an immediate interpretation in terms of SE. Semi-Subsumption is a new and powerful concept, whose nature is still a matter of research, so it seems a worthwhile subject to continue extending this work. On the other hand, the operator approach also provided a new way of using another similar relation: the uncertain conditional. Yet another possible direction of research could be to explore the relation between these two.

Interference Between Occurrences: This work was inspired on an analogy between lexical and ideal quantum measurements and even used another analogy to name the Aboutness Witness, but that is far from exhaust the possible inspiring features of QT for other sciences. Another quantum feature that has been tried for IR is interference, which was quite successfully used for relaxing the independence assumptions of the probabilistic ranking model [71]. Interference also underlies some approaches to IR that try to go beyond bag-of-words and analyse the sequence of terms in a document, like Fourier Domain Scoring [72]. In appendix F we can find a very brief outline of how a complex-valued witness could be defined and used for retrieval. This is still a quite immature idea, but has the potential of giving rise to the formulation of an interesting class of quadratic (instead of linear) witnesses.

7.3.2 Directions for Experimental Research

Inference of Profile Parameters: In chapter 6 a simple way of obtaining information from co-occurrences was used for building the AW. However, in some cases co-occurrence data is quite scarce, and this could limit the performance of a co-occurrence method severely. This suggests the use of more sophisticated inference methods to determine the profiles to build the AW. On the other hand, a consistent tendency was found in the average lexical profiles for terms associated with query terms.

Other IR tasks: The AW can also be used in a range of different IR tasks, and it is probably

better suited for those tasks than it is for *ad hoc* retrieval. In the *filtering* task, for example, the access to documents is restricted, and a stream of documents that are provided to the system, which accumulates information from them at the same time that it has to reject them as non-relevant or accept them relevant [128]. This task would provide the perfect occasion to test the document-defined witnesses ρ that were defined in section 5.4.2 to asses the degree of aboutness between documents.

Appendix A

Dirac Notation

In 1939, for a re-edition of his influencial book *The Principles of Quantum Mechanics* [129], Dirac introduced an elegant notation for vectors and operators on the Hilbert space, that came to be known as **Bra-Ket Notation** or simply **Dirac Notation**. It was thought to bring clarity to the formulation of Quantum Theory, in which it clearly succeeded [130], and is now ubuquitous in physics, especially in the area of Quantum Information and Quantum Computation.

Vectors are described in Dirac notation as a label enclosed between a vertical line and an angle, a graphical remanence of earlier representations as arrows. Vector a is then represented as $|a\rangle$. This representation is called **ket**. Operations like the sum and multiplication by numbers in a field are defined to be associative and distributive:

$$\begin{aligned}(|a\rangle + |b\rangle) + |c\rangle &= |a\rangle + (|b\rangle + |c\rangle) \\ \alpha(\beta|a\rangle) &= (\alpha\beta)|a\rangle\end{aligned} \tag{A.1}$$

$$\alpha(|a\rangle + |b\rangle) = \alpha|a\rangle + \alpha|b\rangle \tag{A.2}$$

Another product can be defined with other vectors: the *inner product*. Since Dirac notation was invented for Quantum Mechanics, where the number field chosen was the complex, an extra concept is defined together with the inner product: the *dual vector*. The dual of a vector is represented as the same label enclosed by the opposite angle and the vertical line, so that the dual of $|a\rangle$ is $\langle a|$. This representation is called **bra**. There is a unary transformation that turns a vector into its dual,

called *Hermitian conjugation*, which is represented by a dagger †:

$$(\alpha|a\rangle)^\dagger = \alpha^*\langle a| \tag{A.3}$$

where $*$ means complex conjugacy.

The dual is defined through another operation: the **inner product**. The inner product of two vectors is a number, and this product is also distributive with respect to the sum and associative with respect to multiplication by a number:

$$|b\rangle \cdot (\alpha|a\rangle) = (|b\rangle)^\dagger(\alpha|a\rangle) = \langle b|(\alpha|a\rangle) = \alpha\langle b|a\rangle \tag{A.4}$$

$$\langle c|(|a\rangle + |b\rangle) = \langle c|a\rangle + \langle c|b\rangle \tag{A.5}$$

Dirac notation was developed to work with numbers in the complex field, so complex conjugation (denoted $*$) has its place in it, closely related to Hermitian conjugation. The complex conjugate of an inner product, is the inner product of the Hermitian Conjugates:

$$(|a\rangle)^\dagger(\langle b|)^\dagger = \langle a|b\rangle = (\langle b|a\rangle)^* \tag{A.6}$$

From relation (A.6) it follows that the inner product of a vector with itself is a real number. If a vector is decomposed in its components of an orthonormal basis $\{|e_i\rangle\}$ where $\langle e_i|e_j\rangle = \delta_{i,j}$, then the inner product with itself will be the sum of the square norm of the coefficients:

$$\begin{aligned} |a\rangle &= \textstyle\sum_i a_i|e_i\rangle \\ \langle a|a\rangle &= \textstyle\sum_i (a_i) * a_i = \sum_i |a_i|^2 \end{aligned} \tag{A.7}$$

The positive square root of the inner product of the vector with itself is called the *norm* of the vector, and is denoted by the label surrounded by double vertical bars:

$$||a|| = \sqrt{\langle a|a\rangle} \tag{A.8}$$

In the same way that the inner product assigns a scalar to every pair of vectors, there is another that assigns a transformation: the *external product*. It is represented as a product of a ket and a bra, in

the order opposite to that of an inner space. How the external product of two vectors transforms a third one, is determined by the inner product:

$$\begin{aligned} T_{b\to a} &= |a\rangle\langle b| \\ T_{b\to a}|c\rangle &= (|a\rangle\langle b|)|c\rangle = |a\rangle\langle b|c\rangle = \langle b|c\rangle|a\rangle \end{aligned} \tag{A.9}$$

Appendix B

Join of Two Rank-One Projectors

Given two vectors $|a\rangle$ and $|b\rangle$, we can decompose the second in a part that is parallel to the first and a part that is orthogonal:

$$|b\rangle = |a\rangle\langle a|b\rangle + \sqrt{1-\langle a|b\rangle\langle b|a\rangle}|\frac{|b\rangle - |a\rangle\langle a|b\rangle}{\sqrt{1-\langle a|b\rangle\langle b|a\rangle}} \tag{B.1}$$

The projector on the minimal subspace containing both vectors (the join subspace) can be obtained by taking the projector on one of the vectors, and adding a projector on the part of the other that is orthogonal to it:

$$|a\rangle\langle a| \cup |b\rangle\langle b| = |a\rangle\langle a| + \frac{(|b\rangle - |a\rangle\langle a|b\rangle)(\langle b| - \langle b|a\rangle\langle a|)}{1-\langle a|b\rangle\langle b|a\rangle} \tag{B.2}$$

It can be easily seen that when the two vectors are both equal (up to an overall phase) or orthogonal, the formula produces the right result. When the two vectors are not the same, a bit of manipulation allows to put this formula as:

$$|a\rangle\langle a| \cup |b\rangle\langle b| = \frac{(|a\rangle\langle a| - |b\rangle\langle b|)^2}{1-\langle a|b\rangle\langle b|a\rangle} \tag{B.3}$$

The symmetric character of this expression (a and b can be permuted with no change in the formula) shows that choosing one or the other vector does not affect the outcome. However, it is worth noting that this formula does not give a definite result when the vectors are equal up to overall phase.

Appendix C

The Meet as a Function of the Sum

The limit of a power of products has been suggested as a means to compute the meet of two projectors in an practical way [61]. However, in some cases, the product of two projectors is not even a hermitian operator, and its power only become one in the infinite exponent limit. Here, another formula is proposed that behaves better. It is always a hermitian operator, even though it is not a projector until the infinite exponent limit. The proposed definition is:

$$A \cap B = \lim_{n\to\infty} \left(\frac{1}{2}(A+B)\right)^n \tag{C.1}$$

The formula has to fulfil two conditions to be considered as a meet:

1. It has to preserve completely any vector that is preserved by both of the individual projectors. This is already fulfiled for $n = 1$:

$$((A|\psi\rangle = |\psi\rangle) \wedge (B|\psi\rangle = |\psi\rangle)) \Rightarrow \left(\frac{1}{2}(A+B)|\psi\rangle = \frac{1}{2}(|\psi\rangle + |\psi\rangle) = |\psi\rangle\right) \tag{C.2}$$

   Factorising $\left(\frac{1}{2}(A+B)\right)^n$ it can be shown that the relation can be obtained from lesser pow-

ers, all the way down to power 1.

$$\left(\frac{1}{2}(A+B)|\psi\rangle = |\psi\rangle\right) \Rightarrow$$
$$\left(\frac{1}{2}(A+B)\right)^n |\psi\rangle = \left(\frac{1}{2}(A+B)\right)^{n-1}\left(\frac{1}{2}(A+B)|\psi\rangle\right) = \left(\frac{1}{2}(A+B)\right)^{n-1}|\psi\rangle$$
$$= \left(\left(\frac{1}{2}(A+B)\right)^n |\psi\rangle = |\psi\rangle\right) \quad \text{(C.3)}$$

2. It has to anihilate any vector that is anihilated by any of the two projectors. Let us supose, for example, that a vector is anihilated by B:

$$\langle\psi|B|\psi\rangle = 0 \quad \text{(C.4)}$$

. In that case, the effect of applying this average of the projectors is:

$$\langle\psi|\frac{1}{2}(A+B)|\psi\rangle = \frac{1}{2}\langle\psi|A|\psi\rangle \quad \text{(C.5)}$$

Since any term where B is applied to the vector becomes zero, Applying the square would result in:

$$\langle\psi|\left(\frac{1}{2}(A+B)\right)^2|\psi\rangle = \left(\frac{1}{2}\right)^2\langle\psi|A(A+B)|\psi\rangle = \left(\frac{1}{2}\right)^2\langle\psi|A^2|\psi\rangle = \left(\frac{1}{2}\right)^2\langle\psi|A|\psi\rangle \quad \text{(C.6)}$$

If we consider higher powers in the same way, all the terms with B will be anihilated, so for power n the result will be:

$$\langle\psi|\left(\frac{1}{2}(A+B)\right)^n|\psi\rangle = \left(\frac{1}{2}\right)^n\langle\psi|A|\psi\rangle \quad \text{(C.7)}$$

The limit of this expression when n tends to infinity, is clearly zero, since $\lim_{n\to\infty}\left(\frac{1}{2}\right)^n = 0$

The join is usually defined from the meet by the de Morgan's law:

$$A \cup B = 1 - ((1-A) \cap (1-B)) \quad \text{(C.8)}$$

where 1 is the identity operator.

However, there is an alternative way of defining the join, also with the sum of projectors, that does not imply a limit:

$$A \cup B = (A + B)(A + B)^{-1} \tag{C.9}$$

where $(A + B)^{-1}$ means the Penrose Generalised Inverse. This function is defined as the operator that fulfils the following relation:

$$XX^{-1}X = X \tag{C.10}$$

Appendix D

Discriminating Products of SEs

Building an optimally discriminating product can be seen as a two-step problem: first, a sequence of central terms is chosen, and then the widths are determined to fulfil the conditions.

The optimal choice of a sequence of central terms can be made by trying every sequence and choosing the best, but the number of possible sequences is astronomical, and this is not a viable procedure. Instead, a greedy search can be done, which starts from a particular term for the SE to be applied first (that on the right end of the product) and chooses the next so as to minimise width, then the next as to minimise widht, and so on, as is explained in the following subsection:

D.0.3 Greedy choice of term sequence

Given a term t_1 to be applied first, a minimum width w^* can be found for any other term t_x such that $|E(t_1, 0)E(t_x, w^*)\ D| > 0$. Then, chose amongst the terms with the lowest w^*. With chosen term t_2, find the minimum width w_1 such that

$$|E(t_2, 0) \circ E(t_1, w_1)\ D| > 0. \tag{D.1}$$

The order-two resulting discriminating product would then be:

$$E(t_2, 0) \circ E(t_1, w_1). \tag{D.2}$$

Then, a minimum width w^* is found for each of the remaining terms t_x, such that

$$E|(t_x, w^*) \circ E(t_2, 0) \circ E(t_1, w_1)\ D| > 0. \tag{D.3}$$

Term t_2 is then chosen amongst those with a lowest w^*, and then minimum width w_2 is chosen such that

$$|E(t_3, 0) \circ E(t_2, w_2) \circ E(t_1, w_1 + w_2)\ D| > 0. \tag{D.4}$$

Then, we have an order-three discriminating product:

$$E(t_3, 0) \circ E(t_2, w_2) \circ E(t_1, w_1 + w_2). \tag{D.5}$$

This way, it is possible to obtain an order-n product:

$$E(t_n, 0) \circ E(t_{n-1}, w_{n-1}) \circ E(t_{n-2}, w_{n-1} + w_{n-2}) \circ \cdots \circ E(t_1, \sum_{i=1}^{n-1} w_i) \tag{D.6}$$

Appendix E

Entanglement and the Entanglement Witness

Entanglement was described by Feynman as **the one and only mystery in Quantum Mechanics** [68] and is generally considered as the most fundamental difference between Quantum and Newtonian mechanics. The possibility of entanglement arises when a physical system is composed of several subsystems. Let us consider two systems A and B, whose states can be represented in Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ respectively. Their joint state would then be represented in the *product* space, denoted $\mathcal{H}_A \otimes \mathcal{H}_B$.

Definition E.1 (Product of Hilbert Spaces)

A **Product of two Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$** is the space spanned by all the tensor products of the orthonormal basis on both spaces $\{|a_i\rangle \otimes |b_j\rangle\}$. If we represent the space itself by the projector on it, the product can be defined as follows:

$$\begin{aligned} \mathcal{H}_A &= \textstyle\bigcup_i |a_i\rangle\langle a_i| \\ \mathcal{H}_B &= \textstyle\bigcup_j |b_j\rangle\langle b_j| \\ \mathcal{H}_A \otimes \mathcal{H}_B = \textstyle\bigcup_{i,j} \left(|a_i\rangle \otimes |b_j\rangle\right)\left(\langle a_i| \otimes \langle b_i|\right) &= \textstyle\bigcup_{i,j} \left(|a_i\rangle\langle a_i|\right) \otimes \left(|b_j\rangle\langle b_i|\right) \end{aligned} \tag{E.1}$$

The concept of product space is quite intuitive: if subsystem A can be in states a_1, a_2 and a_3; subsystem B can be in states b_1 and b_2, then the whole system composed by A and B can clearly be in six states: $(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_2, b_2), (a_3, b_1)$ and (a_3, b_2). A Boolean representation

of these states would be as partitions of a total set, as is shown in figure E.1.

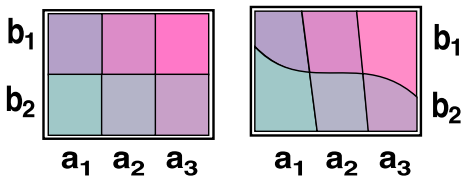

Figure E.1: States of a system composed by two subsystems A and B. The size of each area can represent probability. On the left-hand side, uncorrelated states are found, and on the right, correlated ones (a_1 tends to coincide with b_2, and a_3 tends to coincide with b_1).

When events are independent, their probabilities can be factored out; otherwise, they are said to be *correlated*. A correlation can always be put as a dependence on a common hidden variable:

$$P(a_i, b_j) = \sum_{c_k} P(a_i|c_k)P(b_j|c_k)P(c_k) \neq P(a_i)P(b_i) \tag{E.2}$$

where c_k is a particular value of a hidden variable C that is present in both subsystems with the same distribution.

In Quantum Theory, however, there could be correlated subsystems that cannot be put in terms of local hidden variables; this is called *entanglement*. To define it we need to define a number of classes of states:

Definition E.2 (Uncorrelated State)

An **Uncorrelated State** is a state of a composite system that can be expressed as the product of the states of the subsystems. For a system with N subsystems it would be:

$$\rho_{uncorrelated} = \rho_{subsystem\ 1} \otimes \rho_{subsystem\ 2} \otimes \cdots \otimes \rho_{subsystem\ N} = \prod_{subsystem\ j}^{\otimes} \rho_{(subsystem\ j)} \tag{E.3}$$

The uncorrelated state is totally equivalent to a classical uncorrelated state, where all probabilities are a product of those corresponding to the subsystems.

Definition E.3 (Correlated State)

A **Correlated State** is one that cannot be expressed as a product states, but can be expressed as

a mixture of product states.

$$\rho_{product} = \sum_{product\ i} P_i \prod_{subsystem\ j}^{\otimes} \rho_{(state\ (i,j),subsystem\ j)} \tag{E.4}$$

Correlated states are also equivalent to their classical counterparts, because they can be represented by correlated probability distributions.

Definition E.4 (Entangled State)

An **Entangled State** is one that cannot be expressed as a sum of product states.

$$\forall(\{(\rho_A)_i\},\{(\rho_B)_i\},\{P_i\}), \rho_{entangled} \neq \sum_i P_i\left((\rho_A)_i \otimes (\rho_B)_i\right) \tag{E.5}$$

A dramatic example of entanglement is the so-called Greenberger-Horne-Zeilinger (GHZ) paradox [131]. This consists in a system composed of three subsystems with three binary observables A, B, C each. The whole system is prepared in a superposition state called GHZ state, defined as follows:

$$|ghz\rangle = \frac{1}{\sqrt{2}}\left((|0\rangle_A \otimes |0\rangle_B \otimes |0\rangle_C) + (|1\rangle_A \otimes |1\rangle_B \otimes |1\rangle_C\right) \tag{E.6}$$

If local measurements are made (that is, measurements that only discriminate results of a subsystem) are used, a strange table of results would be obtained. Let us define a set of two binary observables that give an outcome ± 1 for a couple of particular states:

$$\begin{aligned} X &= |0\rangle\langle 1| + |1\rangle\langle 0| = |x_+\rangle\langle x_+| - |x_-\rangle\langle x_-| \\ Y &= i(|0\rangle\langle 1| - |1\rangle\langle 0|) = |y_+\rangle\langle y_+| - |y_-\rangle\langle y_-| \end{aligned} \tag{E.7}$$

where the eigenvalues of these operators are:

$$|x_\pm\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle) \qquad |y_\pm\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm i|1\rangle) \tag{E.8}$$

Supose, now, that we have a device to measure X and two devices to measure Y, for a system in a GHZ state. The results are shown in table E.1.

If we assume that there are simultaneous values for X and Y on every subsystem ready to be measured with whatever correlation, then we could represent them by variables x_A, y_A, x_B, y_B, x_C

A	B	C	result
X	Y	Y	$\langle ghz\|(X \otimes Y \otimes Y)\|ghz\rangle = -1$
Y	X	Y	$\langle ghz\|(Y \otimes X \otimes Y)\|ghz\rangle = -1$
Y	Y	X	$\langle ghz\|(Y \otimes Y \otimes X)\|ghz\rangle = -1$

Table E.1: Results for local measurements with one X and two Y on a GHZ state

and y_B, and we could put the results as the following equations:

$$\begin{aligned} x_A\, y_B\, y_C &= -1 \\ y_A\, x_B\, y_C &= -1 \\ y_A\, y_B\, x_C &= -1 \end{aligned} \tag{E.9}$$

Since every one of these has the value ± 1, then their square is 1, and multiplying the three equations we would obtain:

$$(x_A(y_A)^2)(x_B(y_B)^2)(x_C(y_C)^2) = x_A x_B x_C = -1 \tag{E.10}$$

It seems to be clear that the product of the values of X for the three subsystems would be -1. However, computing this product directly, we obtain:

$$\langle ghz|(X \otimes X \otimes X)|ghz\rangle = 1 \tag{E.11}$$

which is actually the opposite result. Furthermore, there is no assignation of probabilities to values of x and y in the subsystems that can produce these results, no matter how correlated they are. This is known as the GHZ paradox [131]. Quantum Theory solves the paradox, in the sense that it reduces its paradoxical nature to the fact that some properties are taken as not having a defined value.

All this means that GHZ is not either an uncorrelated state, or a correlated state, but something different. To clarify the difference, let us define the quantum concept of *separability*:

Definition E.5 (Separable State)

A quantum state is said to be **separable** if the expected value of every global observable can be expressed as a probabilistic mixture of the product of the expected values of local observables.

$$\exists\{P_k(\{i\}_k)\}, \{|\phi_i\rangle\}),\ \langle\psi| \overset{\otimes}{\prod_i} \hat{O}_i|\psi\rangle = \sum_k P_k(\{i\}_k) \prod_{j\in\{i\}_k} \langle\phi_j|\hat{O}_j|\phi_j\rangle \tag{E.12}$$

In the case of GHZ states, a separate state would be determined by assigning probabilities to every value of x and y for each of the subsystems, that is, assigning $P(x_A, y_A, x_B, y_B, x_C, y_C)$. Since the possible values of these variables are ± 1, there are $2^6 = 64$ possible situations. Some probability distributions over these situations are highly correlated: for example, if $P(+1,+1,+1,+1,+1,+1) = \frac{1}{2} = P(-1,-1,-1,-1,-1,-1)$, there would be no information on the value a variable would take for a particular subsystem, but there would be certainty about all the variables having the same value.

For a given quantum system $|\Psi\rangle$, assigning probabilities to all possible values of internal varables based on the expected value of observables on that state, can be seen as a linear problem:

$$\langle\Psi|\hat{O}_i|\Psi\rangle = \sum_{(x_A,y_A,x_B,y_B,x_C,y_C)} P_\Psi(x_A, y_A, x_B, y_B, x_C, y_C) \cdot O_i(x_A, y_A, x_B, y_B, x_C, y_C) \quad \text{(E.13)}$$

where $\{\hat{O}_i\}$ is the set of measured observables and $O_i(x_A, y_A, x_B, y_B, x_C, y_C)$ is the outcome of its measurement for a given value of the internal variables.

The condition that we can express a particular set of results by assigning probabilities to local values, cuts the space of states in two, and can therefore be represented by an operator which gives negative expected values for states that can be probabilistically described, and positive values for states that cannot. This operator is called *entanglement witness*.

A very important characteristic of the set of entangled states is that **it is not a convex set**. This comes from the fact that there are several possible sets of measurements for which states can produce results impossible to obtain by assigning probabilities to local variables. Each of this set of states would impose a linear restriction that cuts the set of states in two (and therefore define an entanglement witness). However, a separable state must be separable for any possible measurement, so a *family* of entanglement witnesses must be used to assess its separability.

This can be illustrated with a family of GHZ-like states with a varying phase:

$$|GHZ(\theta)\rangle = \frac{1}{\sqrt{2}}\left(|000\rangle + (\cos(\theta) + i * \sin(\theta))|111\rangle\right) \quad \text{(E.14)}$$

The Entanglement Witness for this family corresponding to the choice of measurements (XYY, YXY, YYX and XXX) is:

$$W(\{xyy, yxy, yyx, xxx\}) = |GHZ(\theta = 0)\rangle\langle GHZ(\theta = 0)| - 3|GHZ(\theta = \pi)\rangle\langle GHZ(\theta = \pi)| \tag{E.15}$$

Note that the witness is defined with a state $|GHZ(\theta = 0)\rangle$ for which the value is maximum (1) and other $|GHZ(\theta = \pi)\rangle$ that determines where it starts being negative.

This witness refers only to a particular choice of observables: (XYY, YXY, YYX and XXX). However, for a state to be separable, it should be separable for any set if neasyrents. To show this, we can generate a family of sets of observables by transforming the local observables with an unitary transformation defined localy by:

$$u_\phi = \cos(\phi)\mathbb{I}_{2\times 2} + i\sin(\phi)\sigma_z$$

$$U_\phi = u_\phi \otimes u_\phi \otimes u_\phi \tag{E.16}$$

The witness corresponding to a set of transformed operators $\{U(\phi)O_iU(-\phi)\}$ would be transformed in the same way as the operators:

$$W(\{U_\phi O_i U_{-\phi}\}) = U_\phi W(\{O_i\})U_{-\phi} \tag{E.17}$$

in the particular case defined in (E.15) the transformed witness is:

$$W(\{U_\phi O_i U_{-\phi}\}) = |GHZ(\theta = \psi)\rangle\langle GHZ(\theta = \psi)| - 3|GHZ(\theta = \psi + \pi)\rangle\langle GHZ(\theta = \psi + \pi)| \tag{E.18}$$

In figure E.2 this family of states is represented (in the circumference), together with their probabilistic mixtures (which appear in the area within). An entanglement witness will divide this circle in two with a line, separating a small region of non-separable states and a larger region of separable states. The family of witnesses generated by transforming the observables with U_ψ will also divide the space of states with a straight line, only rotated by ϕ. This is shown in figure E.2. It is important to note that there is a convex set of separable mixed states (represented by the small circle in the middle), but the set of non-separable (entangled) states is non-convex: its representation has a circular hole in the middle.

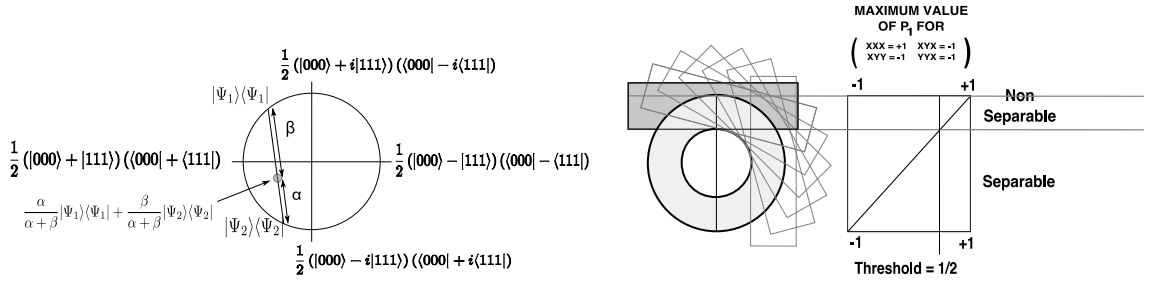

Figure E.2: Parametrised family of GHZ states represented in a circle, Points on the circumference represent pure states, and points within the circle represent mixed states. The grey rectangle shows the area where negative probabilities are necessary to reproduce the results for (XYY, YXY, YYX, XXX). Rotated rectangles show the area where negative probabilities are necessary to reproduce results for observables rotated by a transformation generated by σ_z

Let $|\Psi\rangle$ be a quantum state of a system with N subsystems. We want to check whether $|\Psi\rangle$ admits a probabilistic description, which would mean it is separable. Given a set of local observables $\{\hat{O}_i\}$ acting on the subsystems, we define:

$$V_{i_1,i_2,\cdots,i_s} = \langle\Psi|\hat{O}_{i_1} \otimes \hat{O}_{i_2} \otimes \cdots \otimes \hat{O}_{i_s}|\Psi\rangle \tag{E.19}$$

where s is the number of subsystems. Let us suppose that the value Λ_i of an operator $\hat{O}_i$ depends on internal values of variables x_i associated to a subsystem in a known way:

$$((\langle\phi|X_1|\phi\rangle = x_1) \wedge (\langle\phi|X_2|\phi\rangle = x_2) \wedge \cdots \wedge (\langle\phi|X_s|\phi\rangle = x_s)) \Rightarrow \langle\phi|\hat{O}_i|\phi\rangle = \Lambda_i(\{x_j\}) \tag{E.20}$$

where $|\phi\rangle$ is a state of the subsystem, not to be confused with $|\Psi\rangle$ which is the state of the total system that we want to check for separability.

The problem of finding a probability representation of the system consists in relating these values Λ to the expected values found by computing the probabilities that fulfill the following equation:

$$\sum_{\substack{\text{all values of} \\ \{x_i\}_1, \{x_i\}_2, \cdots, \{x_i\}_s}} \left(\Lambda_{k_1}(\{x_i\}_1)\Lambda_{k_2}(\{x_i\}_2)\cdots\Lambda_{k_s}(\{x_i\}_s) \right) P(\{x_i\}_1, \{x_i\}_2, \cdots, \{x_i\}_s) =$$

$$= \langle\Psi|\hat{O}_{k_1} \otimes \hat{O}_{k_2} \otimes \cdots \otimes \hat{O}_{k_s}|\Psi\rangle \quad \text{(E.21)}$$

This is a simple problem of linear algebra, and can be put in terms of matrices:

$$\begin{pmatrix} M_{(var_1,obs_1)} & M_{(var_2,obs_1)} & \cdots & M_{(var_N,obs_1)} \\ M_{(var_1,obs_2)} & M_{(var_2,obs_2)} & \cdots & M_{(var_N,obs_2)} \\ \vdots & \vdots & \ddots & \vdots \\ M_{(var_1,obs_n)} & M_{(var_2,obs_n)} & \cdots & M_{(var_N,obs_n)} \end{pmatrix} \begin{pmatrix} P_{var_1} \\ P_{var_2} \\ \vdots \\ P_{var_N} \end{pmatrix} = \begin{pmatrix} V_{obs_1} \\ V_{obs_2} \\ \vdots \\ V_{obs_n} \end{pmatrix} \quad \text{(E.22)}$$

or, in Dirac notation,

$$M|P\rangle = |V\rangle \quad \text{(E.23)}$$

where var_i is a set of values for the internal variables $\{\{x_i\}_1, \{x_i\}_2, \cdots, \{x_i\}_s\}$, and obs_j is a set of choices $\{k_1, k_2, \cdots, k_s\}$ of the local operators $\{\hat{O}_k\}$ for all the subsystems. For a given set of internal variables for each subsystem and a given set of operators for the different subsystems, the involved matrices are defined:

$$\begin{aligned} M_{(var,obs)} &= \Lambda_{k_1}(\{x_i\}_1)\Lambda_{k_2}(\{x_i\}_2)\cdots\Lambda_{k_s}(\{x_i\}_s) \\ P_{var} &= P(\{x_i\}_1, \{x_2\}_2, \cdots, \{x_s\}) \\ V_{obs} &= \langle\Psi|\hat{O}_{k_1} \otimes \hat{O}_{k_2} \otimes \cdots \otimes \hat{O}_{k_s}|\Psi\rangle \end{aligned} \quad \text{(E.24)}$$

Since the number of possible values of the internal values is normally much larger than that of possible combinations of observables, equation (E.22) is under-determined, and probabilities cannot be found by inverting matrix M. However, a set of probabilities can be found using Penrose

generalised inverse [132]. Penrose inverse M^{-1_P} of M can be defined so that:

$$\begin{aligned} M \cdot M^{-1_P} \cdot M &= M \\ M^{-1_P} \cdot M \cdot M^{-1_P} &= M^{-1_P} \end{aligned} \tag{E.25}$$

Matrix $M^{-1_P} \cdot M$ is a projector on the space spanned by the rows of M, and we can use it to separate the probabilities vector in a component within that subspace, and the component in the complementary subspace:

$$|P\rangle = \alpha|P_1\rangle + (1-\alpha)|P_2\rangle = (M^{-1_P} \cdot M)|P\rangle + (\mathbb{I}_N - (M^{-1_P} \cdot M))|P\rangle \tag{E.26}$$

It is only $|P_1\rangle$ that can be determined with equation (E.22) just by the formula:

$$\alpha|P_1\rangle = M^{1_P}|V\rangle \tag{E.27}$$

the other part can be obtained from an *a priori* distribution, for example the uniform, by projecting it on the complementary subspace and normalising with norm L^1:

$$|P_2\rangle = \frac{1}{\sum_i \langle i|(\mathbb{I}_N - (M^{-1_P} \cdot M))|P_0\rangle}(\mathbb{I}_N - (M^{-1_P} \cdot M))|P_0\rangle \tag{E.28}$$

where P_0 is an *a priori* probability distribution.

The criterion for separability would be, then, that $|P\rangle = (1-\alpha)|P_1\rangle + \alpha|P_2\rangle$ is a proper probability distribution for some positive value of α. If it has entries that are negative or bigger than 1, it means that the state cannot be described probabilistically in terms of local variables.

E.0.4 An Example

In the case of the GHZ paradox described in section 5.1, two local operators are measured σ_x and σ_y, and six internal variables (x, y for each one of three subsystems) are assigned, combined as $(\sigma_x \otimes \sigma_y \otimes \sigma_y)$, $(\sigma_y \otimes \sigma_x \otimes \sigma_y)$, $(\sigma_y \otimes \sigma_y \otimes \sigma_x)$ and $(\sigma_x \otimes \sigma_x \otimes \sigma_x)$. Given that local variables x and y for each subsystem can only have a value of ± 1, the matrix M of results will also have only values ± 1, coming from products of the local results.

This choice of measurements is neat as an example, because the matrix M of results has a very simple generalised inverse: its transpose, divided by 64:

$$(M_{GHZ})^{-1_P} = \frac{1}{64}(M_{GHZ})^t \qquad M_{GHZ}(M_{GHZ})^t M_{GHZ} = 64 M_{GHZ} \tag{E.29}$$

It is also good because the uniform probability distribution is totally in the complementary subspace:

$$(M^{-1_P}_{GHZ\ paradox} \cdot M_{GHZ\ paradox})|P_{uniform}\rangle = 0|P_{uniform}\rangle \tag{E.30}$$

This means that a probability distribution can be found as:

$$|P(x_A, x_B, x_C, y_A, y_B, y_C)\rangle = \frac{1}{64}(M_{GHZ})^t |V(O_A \otimes O_B \otimes O_C)\rangle + (1-\alpha)|P_{uniform}\rangle \tag{E.31}$$

Since all the entries of $|P_{uniform}\rangle$ are $\frac{1}{64}$, two conditions are then imposed:

1. Lower Bound:
$$\min((M_{GHZ})^t |V(O_A \otimes O_B \otimes O_C)\rangle_i) \geqslant -1 \tag{E.32}$$

2. Upper Bound:
$$\max((M_{GHZ})^t |V(O_A \otimes O_B \otimes O_C)\rangle_i) \leqslant 1 \tag{E.33}$$

GHZ state will produce a vector $((M_{GHZ})^t |V(O_A \otimes O_B \otimes O_C)\rangle$ whose entries are ± 2, clearly violating the conditions for every entry.

A family of parametrised GHZ states will produce different values in the vector $|V\rangle$ of results. In particular, it is interesting to examine the following family with a varying phase:

$$|GHZ(\theta)\rangle = \frac{1}{\sqrt{2}}\left(|000\rangle + (\cos(\theta) + i * \sin(\theta))|111\rangle\right) \tag{E.34}$$

It can be seen that the values of the values vector all vary in a very simple way with the phase,

simply scaled by a factor of $\cos(\theta)$

$$\begin{pmatrix} \langle GHZ(\theta)|(\sigma_x \otimes \sigma_y \otimes \sigma_y)|GHZ(\theta)\rangle \\ \langle GHZ(\theta)|(\sigma_y \otimes \sigma_x \otimes \sigma_y)|GHZ(\theta)\rangle \\ \langle GHZ(\theta)|(\sigma_y \otimes \sigma_y \otimes \sigma_x)|GHZ(\theta)\rangle \\ \langle GHZ(\theta)|(\sigma_x \otimes \sigma_x \otimes \sigma_x)|GHZ(\theta)\rangle \end{pmatrix} = \cos(\theta) \begin{pmatrix} -1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \tag{E.35}$$

With those results, the conditions of positivity become in this case simply $-\frac{1}{2} \leqslant \cos(\theta) \leqslant \frac{1}{2}$. The following operator would then produce a negative expected value for a state that is separable for these measurements, and positive for a state that is not:

$$\begin{aligned} W &= |GHZ(\theta=0)\rangle\langle GHZ(\theta=0)| - 3|GHZ(\theta=\pi)\rangle\langle GHZ(\theta=\pi)| \\ &= (|000\rangle\langle 111| + |111\rangle\langle 000|) - 2(|000\rangle\langle 000| + |111\rangle\langle 111|) \end{aligned} \tag{E.36}$$

Appendix F

Building Witnesses with Complex Coefficients

The final result of a lexical measurement has to be a real number. The use of complex numbers in the representation needs to have a way of hiding imaginary numbers when describing outcomes. Supose, for example, that the lexical profile we want to use is complex-value, because we wont it to include interference between different distances to the central term.

A complex valued witness built with a given set of SEs $\{E_i\}$ can be defined by the value it assigns to a document, as a quadratic function of its the norms of the document transformed by the different SEs:

$$|WD| = \sum_{i,j} \alpha_{i,j} |E_i D| \times |E_j D| \tag{F.1}$$

The coefficients $\alpha_{i,j}$ can be complex, but must comply with a condition:

$$\alpha_{i,j} = (\alpha_{j,i})^* \tag{F.2}$$

where α^* stands for the complex conjugate of α. If $\alpha = a_r + a_i\sqrt{-1}$ then the complex conjugate is the same number with its imaginary part switched sign $\alpha^* = a_r - a_i\sqrt{-1}$. It can also be defined $\alpha^* = \frac{||\alpha||^2}{\alpha}$ where $||\alpha||$ is the complex norm of the number, a generalisation of the absolute value

of real numbers. When $\alpha = a_r + a_i\sqrt{-1}$, the norm is defined $||\alpha|| = \sqrt{(a_r)^2 + (a_i)^2}$

A very simple example of complex AW can be built with a complex profile $\phi_t(w) \neq (\phi_t(w))^*$, as a product of a combination and its conjugate:

$$|WD| = \sum_t \alpha_t \left(\sum_w (\phi_t(w))^* |E(t,w)| \right) \left(\sum_{w'} \phi_t(w') |E(t,w')| \right) \tag{F.3}$$

The resulting value of $|WD|$ will include three different kind of terms:

- diagonal: $|WD|_{diag} = \sum_{t,w} \alpha_t ||\phi_t(w)||^2 |E(t,w)D|^2$
- non-diagonal $|WD|_{non-diag} = \sum_{t,w \neq w'} \alpha_t ((\phi_t(w))^* \phi_t(w') + (\phi_t(w))^* \phi_t(w')) |E(t,w)D| \times |E(t,w')D|$

The first one resembles the usual Witness, but with square counts of nonerased terms (only positive numbers), while the second can inclue negative interference terms between counts at different widths.

Bibliography

[1] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999)

[2] Bellew, R.K.: 1.5: Indexing. In: Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW. Cambridge University Press (2000) 25

[3] Bush, V.: As we may think. ACM SIGPC Notes **1**(4) (1979) 44

[4] Maron, M.E., Kuhns, J.L.: On relevance, probabilistic indexing and information retrieval. J. ACM **7**(3) (1960) 216–244

[5] Bellew, R.K.: 8.2.1: The Finding Out About language game. In: Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW. Cambridge University Press (2000) 299

[6] Jacquemin, C., et, D. In: Term Extraction and Automatic Indexing. Oxford University Press (2000) http://citeseer.ist.psu.edu/jacquemin03term.html.

[7] Cruse, D.: Lexical semantics. Cambridge Univ Pr (1986)

[8] van Rijsbergen, C.J.: A new theoretical framework for information retrieval. In: SIGIR '86: Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1986) 194–200

[9] von Neumann, J.: Mathematical Foundations of Quantum Mechanics. Princeton University Press (1955)

[10] van Rijsbergen, C.J.: The Geometry of Information Retrieval. Cambridge University Press (2004)

[11] Landauer, R.: The physical nature of information. Physics Letters A **217**(4-5) (1996) 188 – 193

[12] Dominich, S.: Mathematical Foundations of Information Retrieval. Kluwer Academic Publishers (2001)

[13] van Rijsbergen, C.: 1: Boolean Models. In: Information Retrieval. Butterworths (1979) http://www.dcs.gla.ac.uk/Keith/Chapter.2/Ch.2.html.

[14] Nie, J.Y., Lepage, F.: Towards a Broader Logic Model for Information Retrieval. In: Information retrieval: uncertainty and logics: advanced models for the representation and retrieval of information. Kluwer Academic Pub (1998) 17–38

[15] Chevallet, J.P., Chiaramella, Y.: Experiences in Information Retrieval Modelling Using Structured Formalism and Modal Logic. In: Information retrieval: uncertainty and logics: advanced models for the representation and retrieval of information. Kluwer Academic Pub (1998) 39–72

[16] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications ACM **18**(11) (1975) 613–620

[17] Widdows, D.: Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In: 41st Annual Meeting of the Association for Computational Linguistics (ACL). (2003) 136–143

[18] Bruza, P., Kitto, K., Nelson, D., McEvoy, C.: Is there something quantum-like about the human mental lexicon? Journal of Mathematical Psychology (2009)

[19] Robertson, S.E., Spärck-Jones, K.: Relevance weighting of search terms. Journal of the American Society for Information Science **27**(3) (1976) 129–146

[20] Nallapati, R.: Discriminative models for information retrieval. In: SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2004) 64–71

[21] Croft, W.B.: Language models for information retrieval. In: 19th International Conference on Data Engineering (ICDE'03), Bangalore, India (2003) 3

[22] Swanson, D.W.: Historical note: Information retrieval and the future of an illusion. Journal of the American Society for Information Science **39**(2) (1988) 92–98

[23] Salton, G., Lesk, M.E.: Computer evaluation of indexing and text processing. Journal of the ACM **15**(1) (1968) 8–36

[24] Webber, S.: Information science in 2003: a critique. Journal of Information Science **29** (2003) 311–330

[25] Kohn, W.: An essay on condensed matter physics in the twentieth century. Reviews in Modern Physics **71** (1999) S59 Example about coexistence of superfluid and normal helium, and how the theory made the experiment possible.

[26] Bellew, R.K.: Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW. Cambridge University Press (2000)

[27] Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. The Kluwer International Series on Information Retrieval. Springer (2005)

[28] Ingwersen, P.: Information Retrieval Interaction. Taylor Graham Publishing (1992)

[29] Saracevic, T.: Relevance: a review of and a framework for the thinking on the notion in information science. (1997) 143–165

[30] Karlgren, J.: Stylistic Experiments for Information Retrieval. PhD thesis, Stockholm University (2000)

[31] Mizzaro, S.: How many relevances in information retrieval? Interacting with computers **10**(3) (1998) 303–320

[32] Hutchins, W.J.: On the problem of 'aboutness' in document analysis. Journal of Informatics **1** (1977) 17–35

[33] Huibers, T., Wondergem, B.: 12: Towards an Axiomatic Aboutness Theory for Information Retrieval. In: Information Retrieval: Uncertainty and Logics - Advanced Models for the Representation and Retrieval of Information. Kluwer Academic Publishers (1998) 297–315

[34] Fang, H., Zhai, C.: An exploration of axiomatic approaches to information retrieval. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2005) 480–487

[35] Huibers, T.W.C., Lalmas, M., van Rijsbergen, C.J.: Information retrieval and situation theory. SIGIR Forum **30**(1) (1996) 11–25 defines Infones, situation, support relation (S |= I). Parallel between indexing and cognition Information more appropriate than Meaning for IR Constraints are relationships between types (P->Q) means (s1 |= P carries the information that s2 |= Q).

[36] Bruza, P.D., Song, D.W., Wong, K.F.: Aboutness from a commonsense perspective. J. Am. Soc. Inf. Sci. **51**(12) (2000) 1090–1105

[37] Luhn, H.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development **1**(4) (1957) 309–317

[38] Lee, L.: Measures of distributional similarity. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics (1999) 25–32

[39] Robertson, S.E., Walker, S., Hancock-Beaulieu, M., Gull, A., Lau, M.: Okapi at trec. In: Text REtrieval Conference. (1992) 21–30

[40] Spärck-Jones, K.: Index terms weighting. Information Storage and Retrieval **9** (1973) 619–633

[41] Singhal, A., Salton, G., Mitra, M., Buckley, C.: Document length normalization. Information Processing & Management **32**(5) (1996) 619–633

[42] Spärck-Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **28** (1972) 11–21

[43] Harter, S.P.: A probabilistic approach to automatic keyword indexing, part i: on the distribution of specialty words in technical literature. Journal of the American Society of Information Science **26** (1975) 280–289

[44] Amati, G., Carpineto, C., Romano, G., Bordoni, F., Roma, I.: FUB at TREC-10 Web Track: A probabilistic framework for topic relevance term weighting. NIST SPECIAL PUBLICATION SP (2002) 182–191

[45] Van Rijsbergen, C.: A theoretical basis for the use of co-occurrence data in information retrieval. Journal of Documentation **33**(2) (1977) 106–119

[46] Wong, S.K.M., Ziarko, W., Wong, P.C.N.: Generalized vector spaces model in information retrieval. In: SIGIR '85: Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1985) 18–25

[47] Cristianini, N., Shawe-Taylor, J., Lodhi, H.: Latent semantic kernels. J. Intell. Inf. Syst. **18**(2-3) (2002) 127–152

[48] Xu, B., Lu, J., Huang, G.: A constrained non-negative matrix factorization in information retrieval. In: IEEE International Conference on Information Reuse and Integration. (2003) 273–277

[49] Hoffman, T.: Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden (1999)

[50] Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical cooccurrence. Behavior Research Methods, Instruments and Computers **28**(2) (1996) 203–208

[51] Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001). (2001) 245–250

[52] Melucci, M.: Exploring a mechanics for context-aware information retrieval. In: Proceedings of the AAAI Spring Symposium on Quantum Interaction, Bremen, Germany, ACM press (2007) 808–815 http://www.dblab.ntua.gr/persdl2007/papers/51.pdf.

[53] Piwowarski, B., Lalmas, M.: A quantum-based model for interactive information retrieval. In: ICTIR. (2009) 224–231

[54] Crestani, F., Pasi, G.: Soft Information Retrieval: Applications of Fuzzy Set Theory and Neural Networks. In: Neuro-Fuzzy Techniques for Intelligent Information Systems. Physica Verlag (Springer Verlag) (1999) 287–315

[55] Cooper, W.S.: Inconsistencies and misnomers in probabilistic ir. In Bookstein, A., Chiaramella, Y., Salton, G., Raghavan, V.V., eds.: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum), ACM (1991) 57–61

[56] Nallapati, R.: Discriminative models for information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM New York, NY, USA (2004) 64–71

[57] Guthrie, L., Walker, E., Guthrie, J.: Document classification by machine: theory and practice. In: Proceedings of the 15th conference on Computational linguistics-Volume 2, Association for Computational Linguistics Morristown, NJ, USA (1994) 1059–1063

[58] Shannon, C., Weaver, W.: A mathematical theory of communication. The Bell System Technical Journal **27**(1928) (July 1948) 379–423

[59] Ponte, J., Croft, W.B.: A language modeling approach to information retrieval. In: Proc. of SIGIR'98, Melbourne, Australia (1998) 275–281

[60] Wei, X., Croft, W.B.: Lda-based document models for ad-hoc retrieval. In: SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2006) 178–185

[61] Beltrametti, E.G., Cassinelli, G.: The logic of Quantum Mechanics. Addison Wesley (1981)

[62] Pfeifer, N., Kleiter, G.: Inference in conditional probability logic. KYBERNETIKA-PRAHA- **42**(4) (2006) 391

[63] Dretske, F.I.: Knowledge and the Flow of Information. Blackwell (1981)

[64] Lalmas, M.: Logical models in information retrieval: Introduction and overview. Information Processing and Management **34**(1) (1998) 19–33

[65] Bennett, J.: A philosophical guide to conditionals. Oxford University Press, USA (2003)

[66] van Rijsbergen, C.J.: A non-classical logic for information retrieval. The Computer Journal **29** (1986) 481–485

[67] G
"ardenfors, P.: Belief revisions and the Ramsey test for conditionals. The Philosophical Review **95**(1) (1986) 81–93

[68] Feynman, R.P.: Lectures on Physics: Quantum Mechanics. Volume 3. Addison-Wesley (1963)

[69] Wooters, W.K.: Quantum mechanics without probability amplitudes. Foundations of Phyics **16**(4) (1986) 391–405

[70] Goyal, P., H., K., Skilling, J.: The origin of complex quantum amplitudes. In Goggans, P.M., ed.: BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: The 29th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Volume 1193., American Institute of Physics (2009) 89–96

[71] Zuccon, G., Azzopardi, L., van Rijsbergen, K.: The quantum probability ranking principle for information retrieval. In: ICTIR. (2009) 232–240

[72] Park, L.A.F., Ramamohanarao, K., Palaniswami, M.: Fourier domain scoring: A novel document ranking method. IEEE Transactions on Knowledge and Data Engineering **16**(5) (May 2004) 529 – 539

[73] Einstein, A., Podolsky, B., Rosen, N.: Can quantum-mechanical description of physical reality be considered complete? Physical review **47**(10) (1935) 777–780

[74] Bell, J.: On the einstein-podolsky-rosen paradox. Physics **1**(3) (1964) 195–200

[75] Bohm, D.: Wholeness and the implicate order. Routledge/Thoemms Press (1981)

[76] Aerts, D., Aerts, S.: Applications of quantum statistics in psychological studies of decision processes. Foundations of Science **1**(1) (1995) 85–97

[77] Busemeyer, J., Wang, Z., Townsend, J.: Quantum dynamics of human decision-making. Journal of Mathematical Psychology **50**(3) (2006) 220–241

[78] Pothos, E.M., Busemeyer, J.R.: A quantum probability explanation for violations of 'rational' decision theory. Proceedings of the Royal Society B: Biological Sciences **276**(1665) (2009) 2171

[79] Khrennikov, A.: Quantum-like formalism for cognitive measurements. Biosystems **70**(3) (2003) 211–233

[80] Conte, E., et al.: A Preliminar Evidence of Quantum Like Behavior in Measurements of Mental States. NeuroQuantol. **6** (2008) 126–139

[81] Schmitt, I.: Qql: A db&ir query language. The VLDB Journal **17**(1) (2008) 39–56

[82] Schmitt, I.: Quantum query processing: unifying database querying and information retrieval. Otto von Guericke Universität Magdeburg, Institut für Technische Informationssysteme (2006)

[83] Muter, P., Marutto, P.: Reading and skimming from computer screens and books: the paperless office revisited? Behaviour and Information Technology **10** (1991) 257–266

[84] Nielsen, Morkes, J., Nielsen, J.: Concise, scannable, and objective: How to write for the web (1997)

[85] Blair, D., Maron, M.: Full-text information retrieval: further analysis and clarification. Information Processing & Management **26**(3) (1990) 437–447

[86] Widdows, D.: A mathematical model for context and word-meaning. In: CONTEXT. (2003) 369–382

[87] Kelvin, B.W.T.L.: Electrical units of measurement. In: Popular Lectures and Addresses. Volume 1., Institution of Civil Engineers, London (May 1889) 73

[88] Scripture, E.W.: The need of psychological training. Science **19** (1892) 127

[89] Carnap, R.: Phylosophical Foundations of Physics. Basic Books (1966)

[90] Buchdahl, G.: Theory cosntruction: The work of norman robert campbell. Isis **55** (1964) 151–162

[91] Campbell, N.R.: An Account on the Principles of Measurement and Calculation. Longmans, Green and Co. LTD. (1928)

[92] Burris, S., Sankappanavar, H.P.: A Course on Universal Algebra. Springer (1981)

[93] Fredkin, E.: Five big questions with pretty simple answers. IBM J. Res. Dev. **48**(1) (2004) 31–45

[94] Michell, J.: Measurement in Psychology: Critical History of a Methodological Concept. Cambridge University Press (1999)

[95] Corbin, J., Strauss, A.: Grounded theory research: Procedures, canons, and evaluative criteria. Qualitative Sociology **13** (1990) 3–21

[96] Beltrametti, E.G., Cassinelli, G.: 11. In: Probability Measures on Orthomodular Posets and Lattices. Volume 15 of Enciclopedia of Mathematics and its Applications. Addison Wesley (1981) 111

[97] Aerts, D.: Quantum interference and superposition in cognition: Development of a theory for the disjunction of concepts (2007)

[98] de Ledesma, L., Phez, A., Borrajo, D., Laita, L.M.: A computational approach to george booleÕs discovery of mathematical logic. Artificial Intelligence **91** (1997) 28 l–307

[99] Boole, G.: An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities. Walton and Maberly (1854)

[100] Hashimoto, J.: On a Lattice with a Valuation. Proceedings of the American Mathematical Society **3**(1) (1952) 1–2

[101] Bell, J., Clifton, R.: Quasiboolean algebras and simultaneously definite properties in quantum mechanics. International Journal of Theoretical Physics **34**(12) (1995) 2409–2421

[102] Rölleke, T., Fuhr, N.: Retrieval of complex objects using a four-valued logic. In: SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1996) 206–214

[103] Lioma, C., Blanco, R., Mochales Palau, R., Moens, M.: A belief model of query difficulty that uses subjective logic. In: International Conference on the Theory of Information Retrieval, Springer (2009) 103

[104] Piron, C.: On the logic of quantum logic. Journal of Philosophical Logic **6**(1) (1977) 481–484

[105] von Neumann, J., Birkhoff, G.: The logic of quantum mechanics. Annals of Mathematics **43** (1936) 298 – 331

[106] Beltrametti, E.G., Cassinelli, G.: 12. In: Characterization of Commutativity. Volume 15 of Enciclopedia of Mathematics and its Applications. Addison Wesley (1981) 125

[107] Hardy, L.: Quantum theory from five reasonable axioms (2001)

[108] Huertas-Rosero, A., Azzopardi, L., van Rijsbergen, C.: Eraser lattices and semantic contents: An exploration of semantic contents in order relations between erasers. In P. D. Bruza, W. Lawless, C.J.v.R., ed.: Proceedings of the III Quantum Interaction Symposium QI2009. Volume 5494 of Lecture Notes in Artificial Inteligence., Springer Verlag (2009) 266–275

[109] Huertas-Rosero, A., Azzopardi, L., van Rijsbergen, C.: Characterising through erasing: A theoretical framework for representing documents inspired by quantum theory. In P. D. Bruza, W. Lawless, C.J.v.R., ed.: Proc. 2nd AAAI Quantum Interaction Symposium, Oxford, U. K., College Publications (2008) 160–163

[110] Katz, S.M.: Distribution of content words and phrases in text and language modelling. Natural Language Engineering **2**(1) (1996) 15–59

[111] Wu, H., Roelleke, T.: Semi-subsumed events: A probabilistic semantics of the bm25 term frequency quantification. In: ICTIR '09: Proceedings of the 2nd International Conference on Theory of Information Retrieval, Berlin, Heidelberg, Springer-Verlag (2009) 375–379

[112] Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: In Proceedings of SIGIR'94, Springer-Verlag (1994) 232–241

[113] Darwin, C.: The Origin of Species by means of Natural Selection. Project Gutemberg (1859)

[114] Beeferman, D., Berger, A., Lafferty, J.: A model of lexical attraction and repulsion. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1997) 373–380

[115] Carpena, P., Bernaola-Galván, P., Hackenberg, M., Coronado, A.V., Oliver, J.L.: Level statistics of words: Finding keywords in literary texts and symbolic sequences. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) **79**(3) (2009) 035102

[116] Ortuño, M., Carpena, P., Bernaola-Galván, P., Muñoz, E., Somoza, A.M.: Keyword detection in natural languages and dna. Europhysic Letters **57** (2002) 759–764

[117] Busch, P.: Quantum states and generalized observables: A simple proof of gleason's theorem. Phys. Rev. Lett. **91**(12) (Sep 2003) 120403

[118] Jozsa, R.: Fidelity for mixed quantum states. Journal of Modern Optics **41**(12) (1994) 2315–2323

[119] Cattaneo, G., Hamhalter, J.: De morgan property for effect algebras of von neumann algebras. Letters in Mathematical Physics **59** (2002) 243–252

[120] Khrennikov, A.: Interpretations of Probability. Walter de Gruyter (2009)

[121] Gleason, A.M.: Measures of the closed subspaces of the hilbert space. Journal of Mathematics and Mechanics **6** (1957) 885–893

[122] Nie, J.Y., Lepage, F.: Information retrieval: uncertainty and logics: advanced models for the representation and retrieval of information. Kluwer Academic Pub (1998)

[123] Kaski, S.: Dimensionality reduction by random mapping: Fast similarity computation for clustering. In: Proceedings of the International Joint Conference on Neural Networks. Volume 1. (1998) 413—418

[124] Wocjan, P., Horodecki, M.: Characterization of combinatorially independent permutation separability criteria. Open Systems & Information Dynamics **12**(4) (2005) 331–345

[125] Shavlik, J., Dietterich, T.: Readings in machine learning. Morgan Kaufmann (1990)

[126] Cristianini, N., Shawe-Taylor, J.: An introduction to support Vector Machines: and other kernel-based learning methods. Cambridge Univ Pr (2000)

[127] Harman, D.K.: Overview of the first text retrieval conference (trec-1). In Harman, D.K., ed.: NIST Special Publication 500-207: The First Text REtrieval Conference (TREC-1). Volume 500 of NIST Special Publications., National Institute of Standards and Technology, NTIS (1992) 1–20

[128] Robertson, S., Hull, D.: The TREC-9 filtering track final report. NIST SPECIAL PUBLICATION SP (2001) 25–40

[129] Dirac, P.A.M.: The Principles of Quantum Mechanics. The International Series of Monographs on Physics. Clarendon Press, Oxford (1930) Series Editors: R. H. Fowler, P. Kapitzer.

[130] Chandrasekhar, S.: PAM Dirac on his Seventieth BirthdayÓ. Contemporary Physics **14** (1973) 389

[131] Greenberger, D.M., Horne, M.A., Shimony, A., Zeilinger, A.: Bell's theorem without inequalities. American Journal of Physics **58**(12) (1990) 1131–1143

[132] Campbell, S.L., C. D. Meyer, J.: Generalized Inverses of Linear Transformations. Dover (1979)