Eraser Lattices for Documents and Sets of Documents

Alvaro Francisco Huertas-Rosero
Uinversity of Glasgow
alvaro@dcs.gla.ac.uk

Abstract

Automatic schemes for the analysis of Natural Language based on word co-occurrence counting have been very successful in capturing meaning, like automatically grouping words referring to similar concepts, or documents about similar topics. In this work, a more general framework is proposed to represent documents and measurements geometrically, in a way directly related with the representation of measurement in Quantum Theory.

Keywords: Natural Language Processing, Quantum Logic, Lexical Measurements, Mathematical Models of Text

1. INTRODUCTION

Co-Occurrence of terms in text has been successfully used for automatically extracting semantic information from text documents (see [3], [4]). In this work, a different approach is proposed, based in transformations that act on documents in a way that is analogous to how projectors act on vectors. These transformations, called Selective Erasers, are defined in section 2. The underlying assumption behind this work is that suitably defined order relations between these measurement transformations are able to capture semantic contents of the text.

2. SELECTIVE ERASERS (SE) AND THEIR INCLUSION RELATIONS

A SE is defined as a transformations that find the occurrences of certain low-level feature in the document, preserve the surroundings, and erase the rest. This general definition is not very useful, because it does not specify what kind of low-level feature can be preserved *together with its surroundings*, and how these surroundings to be preserved can be defined. A more usable definition of a SE is given in [2] for the particular case of term occurrences (as low-level features) in text documents:

> A SE 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.

This concept was first introduced in [1], where some of their properties are shown, in particular, those they share with measurements as described in Quantum Theory. They can also be shown to include well known measurements such as occurrence and co-occurrences of terms and n-grams.

A very important characteristic is that some erasers will *include* others, which means that there will be pairs of Erasers such that what one preserves is included in what the other preserves. Each eraser will preserve small "windows" of text; when those corresponding to eraser A include within them those corresponding to eraser B, we can say that eraser A includes B *for the considered text*. The structure of these relations has been discussed in [2]. The formal condition for an inclusion relation between erasers (which will be denoted $E(t_1, w_1) \succ_D E(t_2, w_2)$ when it holds on document D) would be then:

$$E(t_1, w_1) \succ_D E(t_2, w_2) \iff E(t_2, w_2)E(t_1, w_1)D = E(t_2, w_2)D \quad (1)$$

3. FROM ERASERS TO PROJECTORS

Equation (1) defines a relation that is analogous to the inclusion relation between projectors on subspaces of a vector space. Changing SEs by projectors, and documents by vectors, the relations stand in the same way. The problem of representing Erasers and documents can be addressed through the following ansatz: **For a certain term t, the family of Erasers centred on it $E(t,w)$ would be accurately represented by a set of commuting projectors with rank $f(w)$, where f is a monotonic function**, This way relation $E(t,w) \succ E(t,w+\delta)$ are guaranteed for any integer, positive δ. The correspondence would be:

$$E(t,w) \equiv \Pi_{t,w} = \sum_{i=1}^{f(w)} |\psi_i\rangle\langle\psi_i| \text{ where for any two vectors } |\psi_i\rangle, |\psi_j\rangle \quad \langle\psi_i||\psi_i\rangle = \delta_{(i,j)} \tag{2}$$

Two projectors of the same rank corresponding to different central terms can be converted to each other by a unitary transformation, just like a term-swapping would convert the corresponding SEs:

$$E(A,w) \equiv \Pi_{A,w} = U_{(A \rightleftharpoons B)} \Pi_{B,w} (U_{(A \rightleftharpoons B)})^\dagger = \mathbb{T}_{(A \rightleftharpoons B)} E(B,2) \tag{3}$$

4. RANK OF PROJECTORS

A topic can be thought of as the set of documents about it, and can therefore be represented by inclusion relations. Suppose that for a document D_1 it is the case that $E(A,w_1) \succ_{D_1} E(B,w_2)$, and for document D_2, dealing with the same topic, it holds that $E(A,w_3) \succ_{D_2} E(B,w_4)$. The relation that holds *for both documents* would therefore be descriptive of the topic:

$$\begin{aligned}(E(A,w_1) \succ_{D_1} E(B,w_2)) \ \wedge \ (E(A,w_3) \succ_{D_2} E(B,w_4)) \\ \Rightarrow E(A,\max(w_1,w_3)) \succ_{\{D_1,D_2\}} E(B,\min(w_2,w_4))\end{aligned} \tag{4}$$

The increase in the width difference necessary to produce inclusion relations is crucial to determine the geometric representation of the Erasers. In the example of (4), the difference in width increases from $(w_1 - w_2)$ or $(w_3 - w_4)$ to $(\max(w_1,w_3) - \min(w_2,w_4))$). Empirical evaluations shown in the figure suggest that this width can increase linearly with the number of documents considered.

To draw a vector analogy, we can consider the width factor of a SE can be considered as analogous to the rank of a projector. The join of two projectors will always include both of them, so join projectors can be related in this analogy to an including SE.

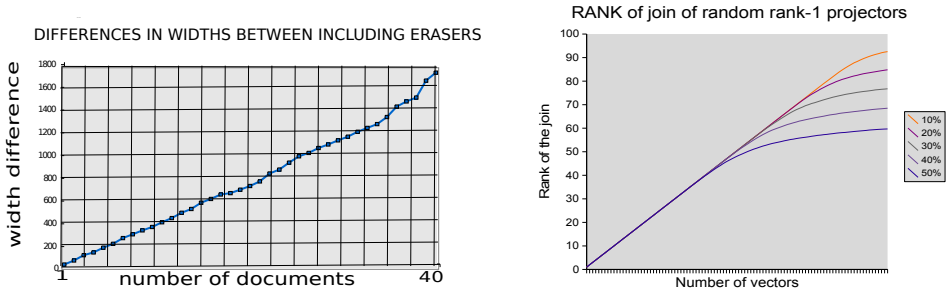

FIGURE 1: Measurement of widths required to produce inclusion relations, on approximately 2000 documents from TREC-1 that were assessed as relevant to 50 different topics. The linear increase of width suggest not to establish direct proportionality between width and rank of the corresponding projector. In the figure on the right, the average rank of joins made with random rank-1 projectors are shown. The different curves represent different threshold criteria to consider a vector as lying in a subspace (threshold ϵ for inner product). The less tight the threshold, the more the line gets closer to the straight line

Let us set a finite threshold for overlap $1 - \epsilon$ to consider a unitary vector as lying in a subspace. Projector Π can be considered as the join of R disjoint 1-dimension subspaces, where R is its trace. A random rank-1 projector will only increase the rank if it is not included in any of these, and since these non-inclusions are independent events, the probability of increasing rank is the product of $R = Tr(\Pi)$ identical terms

$$P(\ Tr(\Pi \cup |\psi\rangle\langle\psi|) = Tr(\Pi) + Tr(|\psi\rangle\langle\psi|)\ ) = (1-\epsilon)^{Tr(\Pi)} \tag{5}$$

The curve showing the expected increase of rank with random vector in a space of dimension 100 is shown in figure 1 suggests that the dimension required to represent erasers as projectors is behind 40. The point where the curve starts showing a negative curvature, like that of the rank curves for projectors, will probably be only approached with bigger collections or more frequent terms. A closer study of this kind of curves could suggest which is the number of dimensions required to represent sets of Selective Erasers as projectors on subspaces of a Hilbert space.

5. ACKNOWLEDGEMENTS

This research has been sponsored by the Department of Computing Sciences of the University of Glasgow, as well as by EPSRC project Renaissance (EP/F014384/1) and Royal Society International Joint Project JP080734. The present work has also benefited from key ideas and guidance from prof. C. J. van Rijsbergen and Dr. Leif Azzopardi, from the Department of Computing Science of the University of Glasgow.

REFERENCES

[1] A.F. Huertas-Rosero, Leif Azzopardi, and C.J. van Rijsbergen. Characterising through erasing: A theoretical framework for representing documents inspired by quantum theory. In C. J. van Rijsbergen P. D. Bruza, W. Lawless, editor, *Proc. 2nd AAAI Quantum Interaction Symposium*, pages 160–163, Oxford, U. K., 2008. College Publications.

[2] A.F. Huertas-Rosero, Leif Azzopardi, and C.J. van Rijsbergen. Eraser lattices and semantic contents: An exploration of semantic contents in order relations between erasers. In C. J. van Rijsbergen P. D. Bruza, W. Lawless, editor, *Proceedings of the III Quantum Interaction Symposium QI2009*, volume 5494 of *Lecture Notes in Artificial Inteligence*, pages 266–275. Springer Verlag, 2009.

[3] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998. http://lsa.colorado.edu/papers/dp1.LSAintro.pdf.

[4] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instruments and Computers*, 28(2):203–208, 1996.