Eraser Lattices and Semantic Contents:
An Exploration of the Semantic Contents in Order Relations between Erasers

Alvaro F. Huertas-Rosero, Leif A. Azzopardi, and C.J. van Rijsbergen

Dept. of Computing Science, University of Glasgow,
Glasgow, United Kingdom
{alvaro,leif,keith}@dcs.gla.ac.uk

Abstract. A novel way to define Quantum like measurements for text is through transformations called Selective Erasers. When applied to text, an Eraser acts like a filter and preserves part of the information of the document (tokens surrounding a central term) and erases the rest. In this paper, we describe how inclusion relations between Erasers can be used to construct an Eraser Lattice for relevant content. It is posited that given a new piece of text, the application of elements of the Eraser Lattice, will result in the destruction or preservation of the content depending on the relevancy of the document. The paper provides the theoretical derivations required to perform such transformations, along with some example applications, before outlining directions and challenges of future work.

1 Introduction

In [1], *Selective Erasers* were proposed as a means for the representation of text documents in a quantum inspired Information Retrieval System. Selective Erasers provide a scheme for lexical measurements in documents, which is analogous to physical measurements on quantum states. In this way, the representation of the text is only known after measurements have been made, and because the process of measuring may destroy parts of the text, the document is characterised through erasure. 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. The count of terms after the transformation is analogous to the formal property of *norm*, and can be represented as such. Given the definition of an Eraser, different lexical measurements can be defined based on it, for example:

1. Occurrence of a term t in document D: $|E(t, 0)D|$
2. Frequency of occurrence of a term t in document D: $\frac{|E(t,0)D|}{|D|}$
3. Co-occurrence frequency of terms t_1 and t_2 in document D with a minimum distance w: $\frac{|E(t_2,0)E(t_1,w)D|}{|D|}$

P. Bruza et al. (Eds.): QI 2009, LNAI 5494, pp. 266–275, 2009.

where |.| is a counting operation. While this constitutes a basis of representation of text documents, a method is required in order to harness the analogy, that is, to perform some higher level retrieval operation. In this paper, we extend the formalisation of Selective Erasers to Selective Lattices, which are used to performing ranking or classification based on the "quantum" representation of documents. We posit that it is possible to define a set of compatible Erasers which characterise relevancy, such that the application of these Erasers will either preserve a document or destroy it. If a document is preserved (or largely preserved) then this is indicative of its relevance, while if a document is destroyed (or largely destroyed) then this is indicative of its non-relevance. Specifically, we hypothesise that:

> for a given query, the relations between a set of optimally chosen Erasers will differ significantly in the subset of relevant documents and in the subset of the non-relevant documents.

Thus, we believe that we can characterise the relevancy and non-relevancy through erasure. The intuition is that the usage of language within relevant documents will be similar and that the erasers will preserve this usage, while in non relevant documents the usage of language will be different, even if the same vocabulary is used, and thus be erased.

The remainder of this paper will be as follows: The next section will defined the necessary order relations between erasers, i.e. strict ordering and orthogonality. Section 3, will describe how the Eraser Lattice can be constructed using a partially ordered set, before describing how to use the Eraser Lattice to classify documents as either relevant or non relevant. Then, in Section 5, we perform an empirical study on a standard IR test collection (AP88) where we demonstrate the utility of the method and show how relevance information can be preserved through optimally selected Erasers. Finally, we conclude with a discussion of this work and directions for future work.

2 Erasers and Their Order Relations

As a strategy to catch the context in which words tend to occur, in this work we propose to examine relations between Selective Erasers associated with the occurrence of different terms. Several relations can be defined between Selective Srasers as acting on a certain document, but in this work we focus on two of them, orthogonality and strict ordering (others are also mentioned in appendix A):

- **Orthogonality (Disjointedness):** Two Erasers are orthogonal when there is no common fractions of a document D they both preserve:

$$E_1 \perp_D E_2 \iff \forall D_i \;\; |E_1[E_2 D_i]| = 0 \tag{1}$$

- **Strict Ordering (Inclusion):** An order relation exists when one Eraser includes the other, that is, when everything one Eraser preserves in document

D, the other preserves as well. A formal way of stating it for two Erasers E_1 and E_2 is that defined for projectors in [2]:

$$E_1 \geqslant_D E_2 \iff \forall D_i \;\; E_2\,[E_1 D_i] = E_2 D_i \tag{2}$$

These relations could also be defined within a subset of the documents, when they hold *for every document in subset s*. This relation within a subset is represented with a subscript on the symbol of the relation. For example, for strict inclusion, it would be

$$E_1 \geqslant_S E_2 \iff \forall D_i \in S \;\; E_2\,[E_1 D_i] = E_2 D_i \tag{3}$$

The number of possible Erasers for all the vocabulary in a collection is astronomical, so the practical applicability of this criterion relies on a sensible scheme for selecting Erasers and relations between them. Our approach to that problem is based on the measurement of extremal (maximal or minimal) distances between occurrences of terms.

2.1 Distances between Occurrences and Order Relations

Let us suppose that terms t_1 and t_2 occur in document D $n_1 \neq 0$ times and $n_2 \neq 0$ times respectively. 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, two nontrivial relations can be defined that are fulfilled within this document:

$$E(t_1, d_{min} - \delta_1) \quad \perp_D \quad E(t_2, \delta_1) \tag{4}$$

$$E(t_1, d_{max(t_1,t_2)} + \delta_2) \quad \geqslant_D \quad E(t_2, \delta_2 - \delta_3) \tag{5}$$

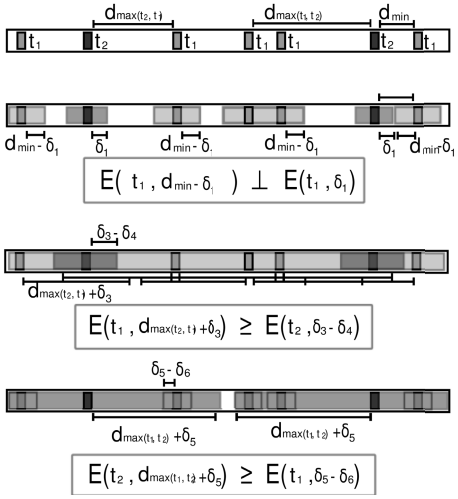

Fig. 1. Relations between Erasers for maximum and minimum distances between occurrences

where δ_1, δ_2 and δ_3 are natural numbers that can vary freely (as long as the width factor remains equal or bigger than zero). Extremal distances show how wide or narrow an Eraser have to be to include or avoid another, as is illustrated in figure 1. The difference between $d_{max(t_1,t_2)}$ and $d_{max(t_2,t_1)}$ is also explicit in the figure.

3 Computation of Eraser Lattices

Any set of Erasers forms, with their order relations, a *Partially Ordered Set* (poset), since relation $\geqslant$ is a proper order relation (reflexive, antisymmetric and transitive). It is not a totally ordered set because there are not order relations between every pair of Erasers [3]. Furthermore, to make it a *lattice* it is necessary to augment it with an *infimum*, a transformation that erases everything, and a *supremum*, a transformation that does not erase anything.

A lattice can be represented by a Hasse diagram, where the infimum is below, the supremum is above, and the elements are in the middle connected with vertical or diagonal lines whenever an order relation holds. In figure 2 the lattice corresponding to the example in figure 1 is depicted.

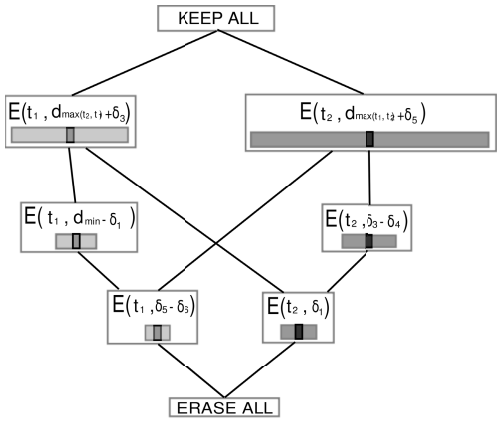

Fig. 2. Hasse diagram representing a lattice

What is *Quantum* About This Scheme? A close inspection of figure 2 can reveal a very interesting feature of the lattices of Erasers: they are non-distributive. Missing crossed relations between the four Erasers below the upper point (supremum) produce 4 possible sub-lattices with the shape of a pentagon called N_5 (for the pentagon, see [3], and for non-distributive logics, [4]) which are known to be a signature of non-boolean lattices. This is not just a mathematical curiosity in the structure of order relations: it means that usual boolean relations are restricted to hold *within* sets of compatible measurements, but they will not hold between elements of two different (incompatible) sets. In Quantum

Theory, sets of compatible observations are related to a particular experimental or operational context; thus in this work, we suggest that sets of compatible lexical measurements could also be related to some kind of context. In particular, we explore the possibility to relate them to a *topical* context, therefore using the contextual nature of Quantum Logics to explore topicality.

Relevancy-Sensitive Eraser Lattices. Relations between Erasers are determined in most cases by the usage of terms. A set of Erasers can be chosen, that when applied to a set of relevant documents, possesses a lattice structure. It is our hypothesis that such lattice would encode semantic information about the topic the documents in the set are relevant to, and it could be used to define a transformation that preserves as much as possible of relevant documents and as least as possible of nonrelevant documents (a *Topical Eraser*).

The first approximation to a Topical Eraser is through a set of orthogonal (disjoint) Erasers. Orthogonality tends to enforce a low window width (Erasers with a width factor of 0 are all orthogonal to each other), so it is desirable to choose them with maximally wide windows, to enhance document preservation in the set of relevant documents. Thus, a set of orthogonal Erasers with *maximum window width* are chosen for the set of relevant documents, and the Topical Eraser can be defined as one that preserves what any of these preserve, that is, their join (union):

$$E_{topic} = \bigcup_i E(t_i, n_i) \tag{6}$$

The fraction X of a relevant document $D \in topic$ preserved by this Topical Eraser would be extremely easy to compute. Since they are orthogonal, the fraction preserved by the join would be simply the sum of the individual preserved fractions. And this fraction, in turn, would be approximately proportional to the occurrence frequency of the terms, except for border effects (windows truncated by the beginning or end of the document):

$$X_{topic}(D) = \frac{|E_{topic}D|}{|D|} = \frac{|(\bigcup_i E(t_i, n_i))D|}{|D|} = \sum_i \frac{|E(t_i, n_i)D|}{|D|} \approx \sum_i (2n_i + 1)F(t_i) \tag{7}$$

where $F(t_i)$ is the frequency of occurrence of term t_i (the occurrence divided by the length of the document) and n_i is the window width parameter. This way, we get something like a TF (Term Frequency) scoring with occurrence of terms and weighting factors (widths) tuned with the set of Relevant Documents.

In a non-relevant document, on the other hand, the sum expression would not be valid, since the Erasers would not be necessarily orthogonal, and the fraction preserved by the join would be less than the sum of the fractions preserved by the individual Erasers (since terms in overlaps are not counted twice).

The terms chosen for this set could still occur frequently in nonrelevant documents, producing high *preserved fraction* in nonrelevant documents, and therefore poor sensitivity to context. To avoid this, a further set of Erasers can be used. With the data of *maximal* distances between occurrences, a set of Erasers

can be defined such that each one of them *includes* one of the previous Erasers. Since inclusion relation (5) tends to favour Erasers with wide windows, those with *minimal window width* can be chosen to enhance sensitivity. For ED_i being one of the chosen maximal disjoint Erasers, and EI_i the corresponding including Eraser, the condition would be:

$$EI_i = E(t_j, n_j) \qquad \text{such that} \qquad (EI_i \geqslant ED_i) \wedge (n_{k\neq j} > n_j) \tag{8}$$

On a relevant document, the consecutive application of the including Eraser and the disjoint Eraser would produce the same result than just the application of the disjoint Eraser, but would erase more than the disjoint Eraser in a nonrelevant document, where the inclusion relation (8) does not necessarily hold.

4 Choice of Erasers

Central Terms: To choose the central terms for the Erasers, the ratio between the *average distance to occurrence of other terms* and the *average distance to occurrences of itself* can be used as a criterion to choose terms. Terms would be ranked according to the following quantity:

$$R(t_i) = \begin{cases} \frac{\langle d_{(t_i,t_i)} \rangle}{\langle d_{(t_i,t_{j\neq i})} \rangle} & \text{when present} \\ 0 & \text{when absent} \end{cases} \tag{9}$$

where $\langle \cdot \rangle$ means average and $d_{(t_i,t_j)}$ is the distance between an occurrence of t_i and the nearest occurrence of t_j. If the term is absent in a document, this would count as a 0 in the averaging.

A term that tends to be evenly spaced in the text and occur relatively near to everyone of the others would score high, and one that either occurs very concentrated or does not occur much, will get a low score.

Window Widths: There are two possible criterion that we can use to assign window widths both in disjoint and including Erasers:

1. Maximum *preserved fraction*: This criterion favours maximal window widths for disjoint Erasers
2. Minimum overlap: This criterion favours minimal window widths for including Erasers.

In different documents, the maximum widths compliant to orthogonality condition (4) and the minimum widths compliant to inclusion condition (5) can be different, so we maximise or minimise them, correspondingly, over the whole set of documents. Minimum distance will then be minimum in all the set of documents, and maxima will be also maxima on all the set.

5 A Practical Example in Collection AP88

To check to what extent semantic contents is encoded in the order relations between Erasers, we chose 2 sets of 20 Erasers for the set of relevant documents

Table 1. Erasers chosen for query 82 of AP88

Query 82: Genetic Engineering			
term disjoint	width disjoint	term including	width including
said	13	field	81
field	59	corn	36
new	13	said	25
year	46	genetically	144
s	22	scientists	144
t	33	s	36
used	34	test	81
research	30	t	121
tests	50	aids	81
test	55	disease	100
researchers	31	genetic	100
disease	46	research	100
gene	40	used	81
cancer	37	gene	81
scientists	21	researchers	100
corn	61	vaccine	36
aids	47	cancer	36
genetically	45	new	64
vaccine	46	tests	64
genetic	22	year	121

for different topics in the collection AP88, and for each topic compared the relations holding in the set of *relevant* documents and those holding in a random subset of *nonrelevant* documents.

Results are in table 1 for topic 82. Central terms are clearly related to the topic. The fulfilment of an order relation can be approximately evaluated by comparing the documents acted upon by both Erasers and only one of them, as follows:

$$X(E_1 \geqslant_D E_2) = \text{sim}(E_2 E_1 D, E_2 D) \tag{10}$$

where $\text{sim}(A, B)$ is a measurement of the similarity of documents A and B. In figure 3 a part of the lattice for topic 58 is depicted. The most important test for this scheme is the measure of the discrimination between relevant and

Table 2. Topics that were well characterised (easy) and poorly characterised (difficult). The average number of documents and percentages of preservation of different kinds of documents are presented. These values correspond to Topical Erasers with 20 central terms.

Queries	documents	% relevants	% nonrelevants	% nonasessed
easy queries	42.7 ± 26.87	(72.36 ± 8.81)%	(30.47 ± 12.87)%	(9.64 ± 7.23)%
difficult queries	93.69 ± 33.15	(46.62 ± 8.33)%	(70.31 ± 6.06)%	(11.15 ± 9.48)%

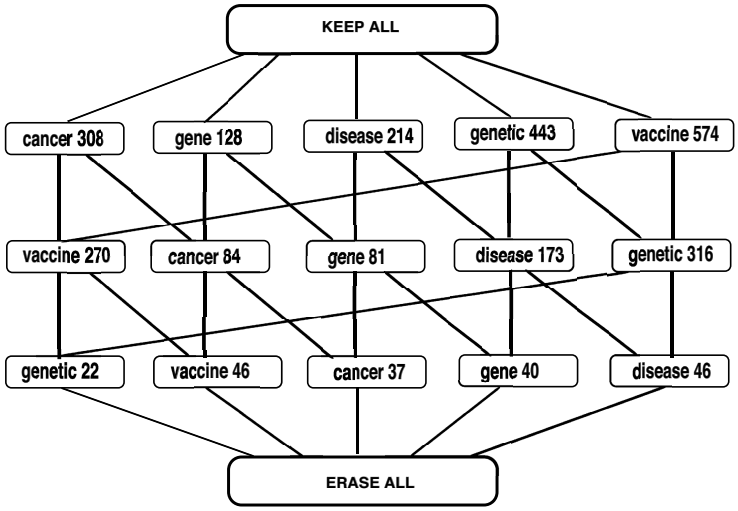

Fig. 3. Order relations between Erasers for topic 82 of TREC-1. Relations are obtained from the list of including Erasers with formulas in appendix B.

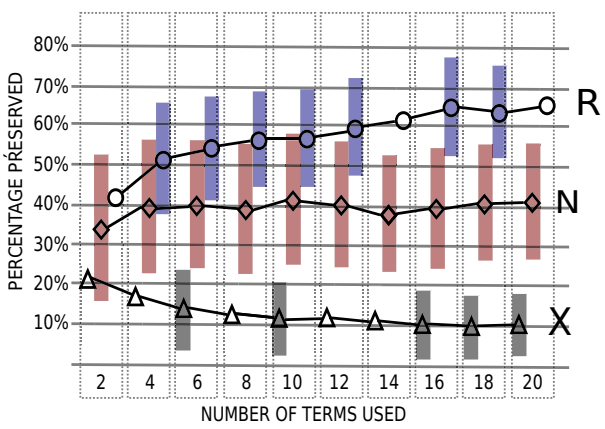

Fig. 4. Average preserved percentage of relevant (R), non-relevant (N) and non-assessed (X) documents for queries 51-100 of TREC-1, for different numbers of central terms

nonrelevant documents. In figure 4 the results are shown for the percentages preserved for queries 51 to 100 of TREC-1 [5] using different numbers of central terms to build the Topical Eraser. Relevant documents were well characterised by the Topical Eraser for most of the queries, but 13 of them had bigger preserved percentage for nonrelevant than for relevant documents. In table 2, results of preserved percentage are shown for the topics (queries) in two groups: the 13

queries for which the results were anomalous (more non-relevant than relevant preserved) and those for which the methodology worked as expected, preserving more relevant than non-relevant.

6 Conclusions

In this paper, we have extended the notion of Selective Erasers to form higher order constructs called selective lattices. Constructing lattices from a set of example relevant documents is a novel way in which to capture the semantic relations within the content. The application of transformations derived from elements of the lattice, will either (mostly) destroy or preserve the relevant information in a new unseen document, and provides a formal mechanism for classifying documents. Examples with sets of documents of a standard IR test collection were used to check the ability of this scheme to capture semantic contents, with positive results.

Future work will be directed towards deriving the optimal set of Selective Erasers, formulating a ranking algorithm and performing a large scale empirical study of one of the first quantum inspired Information Retrieval System. It could also be possible to check situations where incompatible observations exist, like a possible incompatibility between topicality and relevancy mentioned in [4, chapter 6].

Acknowledgements

We would like to thank Guido Zuccon for his valuable input and suggestions. This work was sponsored by the European Comission under the contract FP6-027026 K-Space and Foundation for the Future of Colombia COLFUTURO.

References

1. Huertas-Rosero, A., Azzopardi, L., van Rijsbergen, C.: Characterising through erasing: A theoretical framework for representing documents inspired by quantum theory. In: Bruza, P.D., Lawless, W., van Rijsbergen, C.J. (eds.) Proc. 2nd AAAI Quantum Interaction Symposium, Oxford, U. K., pp. 160–163. College Publications (2008)
2. Beltrametti, E.G., Cassinelli, G.: 9. In: The logic of Quantum Mechanics, p. 87. Addison-Wesley, Reading (1981)
3. Burris, S., Sankappanavar, H.P.: A Course on Universal Algebra. Springer, Heidelberg (1981)
4. van Rijsbergen, C.J.: The Geometry of Information Retrieval. Cambridge University Press, Cambridge (2004)
5. 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). NIST Special Publications, vol. 500, pp. 1–20. National Institute of Standards and Technology, NTIS (1992)

A More Relations between Erasers

1. Trace Ordering (does not rely on the identity between transformed documents but only between their preserved number of tokens).

$$E_1 \geqslant_{(st,D)} E_2 \iff |E_2E_1D| = |E_2D| \tag{11}$$

2. Weak Trace Ordering (this is rather trivial, and not very useful):

$$E_1 \geqslant_{(wt,D)} E_2 \iff |E_1D| = |E_2D| \tag{12}$$

   It can be easily shown that the three defined relations form a chain of implication (3) ⇒ (13) ⇒ (12).
3. Compatibility

$$E_1 \sim_D E_2 \iff [E_2E_1]D = [E_2E_1]D \tag{13}$$

   This relation implies all the other relations defined in this paper, but is not necessary.

B Deducing More Relations

Two relations are very useful to deduce more order relations from a list, like that of narrow-window Erasers and wide-window Erasers:

1. Transitivity:

$$(E(A, w_A) \geqslant E(B, w_B)) \wedge (E(B, w_B) \geqslant E(C, w_C)) \Rightarrow (E(A, w_A) \geqslant E(C, w_C)) \tag{14}$$

2. Invariance under simultaneous widening:

$$\forall \alpha > 0, (E(A, w_A) \geqslant E(B, w_B)) \Rightarrow (E(A, (w_A + \alpha)) \geqslant E(B, (w_B + \alpha))) \tag{15}$$