# Negations and document length in logical retrieval

David E. Losada[1] and Alvaro Barreiro[2]

[1] Intelligent Systems Group,
Department of Electronics and Computer Science,
University of Santiago de Compostela, SPAIN
dlosada@dec.usc.es
[2] AIlab,
Department of Computer Science,
University of A Coruña, SPAIN
barreiro@udc.es

**Abstract.** Terms which are not explicitly mentioned in the text of a document receive often a minor role in current retrieval systems. In this work we connect the management of such terms with the ability of the retrieval model to handle partial representations. A simple logical indexing process capable of expressing negated terms and omitting some other terms in the representation of a document was designed. Partial representations of documents can be built taking into account document length and global term distribution. A propositional model of information retrieval is used to exemplify the advantages from such expressive modeling. A number of experiments applying these partial representations are reported. The benefits of the expressive framework became apparent in the evaluation.

## 1 Introduction

For many retrieval systems the set of terms that determines the rank of a certain document given a query is solely composed of the terms in common between document and query. Nevertheless, it is well known that documents are often vague, imprecise and lots of relevant terms are not mentioned. Since topicality is a key component of retrieval engines, models of Information Retrieval (IR) should avoid to take strong decisions about the relationship between terms and document's semantics.

Current practice in IR tends to limit unfairly the impact of terms which are not explicitly mentioned by a given document. Although the vector-space model maintains a dimension for every term of the vocabulary, popular weighting schemes assign a null weight for those terms not explicitly mentioned. Similarly, probabilistic approaches, whose basic foundations allow to consider all the terms of the alphabet to do retrieval, tend to reduce the computation to the set of terms explicitly mentioned by a given document [14]. A notable exception is located in the context of the Language Modeling (LM) approaches [9, 2]: a term $t$ which is not present in a document $d$ is not considered as impossible in connection with the document's semantics but $t$ receives a probability value greater than zero. This value grows with the global distribution of the term in the document collection, i.e. if $t$ is frequently used by documents in the collection then it is possibly related to the document $d$. This is a valuable approach because it opens a new way to handle terms not explicitly mentioned in a given document but, on the other hand, the opposite problem arises: no one term can be considered totally unrelated to a given document. This is because all the probability values coming from every query term are multiplied together and, hence, if zero probabilities are allowed then we would assign a null probability to any document that regards one or more of the query terms as unrelated.

In this work we propose an alternative way for handling both situations. A term $t$ which is not explicitly mentioned by a document $d$ may be considered as: a) totally unrelated to $d$ and, hence, if a query uses $t$ then the document $d$ is penalized (this penalization should not be as extreme as assigning a retrieval status value of 0 for $d$) or b) possibly related to $d$ and, hence, a non-zero contribution is computed for modeling the possible connection between $t$ and $d$.

A formalism allowing partiality can distinguish between: a) lack of information about the actual connection between a given topic and a particular document, b) certainty that a given topic is completely out of the scope of a given document and c) certainty that a given topic is totally connected to the contents of a given document. In particular, logic-based models [15, 1] supply expressive representations in which these situations can be adequately separated. In this work we use a logical model of IR based on Propositional Logic and Belief Revision (PLBR) [6, 8] to exemplify the advantages of the logical modeling. We design a novel logical indexing method which builds expressive document representations. The logical indexing is driven by global term distribution and document length. In this way, intuitions applied in the context of document length normalization [13, 11] and LM smoothing techniques [9] can be incorporated into the logical formalism. This indexing approach was empirically evaluated revealing the advantages of the approach taken.

The rest of this paper is organized as follows. In section 2 the foundations of the logical model are presented. This section is intentionally brief because further details can be found in the literature. Section 3 addresses the construction of partial representations for documents in connection with global term distribution and document length. Experiments are reported in section 4 and section 5 offers an analysis a posteriori of the behaviour of the indexing method. Some conclusions and possible avenues of further research are presented in section 6.

## 2 The model

Given a document and a query represented as propositional formulas $d$ and $q$, respectively, it is well known that the notion of logical consequence (i.e. $d \models q$) is rather strict for retrieval because it yields a binary relevance decision [15]. The PLBR model defines a measure of closeness between $d$ and $q$ which can straightforwardly be used to build a formal rank of documents induced by the query [6, 8].

Dalal's Belief Revision measure of distance between logical interpretations [3] stands on the basis of the PLBR approach. A query $q$ can be seen as the set of logical interpretations satisfying $q$, i.e. the set of *models* of $q$. The distance from each model of the document to the query is computed as the minimum distance from the model of the document to the query models. The final distance from the document to the query is the average distance from document models to the query.

Given a model of the document and a model of the query, the original PLBR distance basically counts the number of *disagreements*, i.e. the number of propositional letters with different interpretation[3]. This approach was later extend to define a new measure of distance between logical interpretations that takes into account inverse document frequency (idf) information [4]. Within this measure, every letter mapped into the same truth value by both interpretations produces an increment to the final distance that depends on its idf value. Note that this extension maintains the propositional formalism for representing documents and queries but introduces idf information for distance computation. As it will be explained later, in this paper we will use idf information for producing negated terms in the logical document representations. Observe that both uses of idf information are different because the former is done at matching time whereas the latter is done at indexing time.

The PLBR distance can be computed in polynomial time provided that $d$ and $q$ are in disjunctive normal form (DNF) [7]. A prototype logical system was implemented to evaluate the PLBR model against large collections. The experiments conducted revealed important benefits when handling expressions involving both logical conjunctions and disjunctions [5].

Nevertheless, the logical indexing applied so far was rather simplistic. No major attention was paid to the design of evolved techniques to produce more expressive document logical representations. In particular, the use of logical negations was left aside, which is precisely the aim of this work.

## 3 Partial representations for documents

The PLBR model has provision for establishing a distinction between a term for which we do not know whether or not it is significant with respect to a given document's semantics and a term for which we have positive evidence that it is not related at all with document's contents. The latter case naturally leads to a negated expression of the term within document's representation whereas the most sensible decision regarding the former case is to omit the term in the document's representation.

We first present some heuristics that can be applied to identify appropriate terms to be negated and then we give a further step to connect the new logical indexing with document's length.

### 3.1 Negative term selection

Let us consider a logical representation of a document in which only the terms that appear in the text of the document are present as positive literals[4] (let us call this conservative setting as *negate-nothing approach*). Of course, many of the terms not mentioned by the document (*unseen* terms) will undoubtedly be disconnected with document's

[3] Note that every indexing term is modelled as a propositional letter in the alphabet.

[4] A literal is a propositional letter or its negation.

contents and, hence, to omit those terms within the document's representation does not seem to be the best choice. On the contrary, a negated representation of those terms appears as a good alternative. To negate every unseen term is also unfair (*negate-all approach*) because there will be many topics that, although not explicitly mentioned, are strongly connected with document's semantics.

We propose a logical indexing strategy that negates *some* unseen terms selected on the basis of their distribution in the whole collection. Note that this global information is also used in the context of LM smoothing strategies for quantifying the relatedness between unseen terms and document's contents. More specifically, a null probability is not assigned for a term which was not seen in the text of a document. The fact that we have not seen it does not make it impossible. It is often assumed that a non-occurring term is possible, but no more likely than what would be expected by chance in the collection.

If a given term is infrequent in the document base then it is very unlikely that documents that do not mention it are actually related to this topic (and, thus, very unlikely that any user that wants to retrieve those documents finds the term useful when expressing her/his information need). On the other hand, frequent terms are more generic and have more chance to present connections with the topics of documents even in the case when they are not explicitly mentioned. This suggests that unseen infrequent terms are good candidates to formulate negations in the logical indexing process.

The obvious intention when negating a term in a document's representation is to move the document away from queries mentioning the term. Consider a query term which is missing in the text of a given document. If the query term is globally infrequent and, thus, it had been negated within the document's representation then the document will be penalized. On the contrary, if the term is globally frequent and it was omitted in the document's representation, then the penalization is much lower. This is intuitive because frequent terms have much more chance of being connected with the contents of documents that do not explicitly mentioned them.

### 3.2 Document length

We now pay attention to the issue of the number of terms that should be negated in the representation of every document. In this respect, a first question arises: is it fair to negate the same number of terms for all documents? In the following we try to give a motivated answer.

Let $T_d$ be the subset of terms of the alphabet ($T$) that are present in the text of a document $d$. Consider that we decide to introduce $k$ negated terms in the logical representation of $d$. That is, every term in $T_d$ will form a positive literal and $k$ terms in $T \backslash T_d$ (the $k$ terms in $T \backslash T_d$ most infrequent in the collection) produce $k$ negated literals. If we introduce the same number of negations for all documents in the collection we would be implicitly assuming that all documents had the same chance of mentioning explicitly all their relevant topics. This assumption is not appropriate.

A long document may simply cover more material than a short one. We can even think on a long document as a sequence of unrelated short documents concatenated together. This view is called the *scope hypothesis* and contrasts with the *verbosity hypothesis*, in which a long document is supposed to cover a similar scope than a short document but simply uses more words [11]. It is accepted that the verbosity hypothesis prevails over the scope hypothesis. Indeed, the control of verbosity stands behind the success of high performance document length normalization techniques [13, 11].

This also connects with recent advances on smoothing strategies for Language Modeling. For instance, a bayesian predictive smoothing approach takes into account the difference of data uncertainty in short and long documents [16]. As documents are larger, the uncertainty in the estimations becomes narrower. A similar idea will drive our logical indexing process because long documents are supposed to indicate more exhaustively their contents and, hence, more assumptions on the non-related terms will be taken.

A fixed number of negations for every document is also not advisable from a practical perspective. Think that the sets $T \setminus T_d$ are very large (because $T_d << T$) and, hence, there will surely be many commom terms in the sets $T \setminus T_d$ across all documents. As a consequence, there will likely be little difference between the negated terms introduced by two different documents and, therefore, the effect on retrieval performance will be unnoticeable.

In this work we propose and evaluate a simple strategy in which the number of negations grows linearly with the size of the document. In our logical indexing process, the size of a document will be measured as the number of different terms mentioned by the document.

Another important issue affects the maximum and minimum number of negations that the logical indexing will apply. Let us assume that, for a given document $d$, we decide to include $1000$ negated literals in its logical representation. Since the number of negations is relatively low (w.r.t. current term spaces), the involved terms will be very infrequent, most of them mentioned by a single document in the whole collection and, therefore, it is also very unlikely that any query finds them useful to express an information need. As a consequence, a low number of negations will definitely not produce any effect on retrieval performance because the negated terms are rare and will be hardly used by any query. This advances that significant changes on the retrieval behaviour of the logical model will be found when the number of negations is high. Inspired by this, we designed our logical indexing technique starting from a total closed-world assumption (i.e. we negate every unseen term) and we reduce the number of negations as document's size decreases. That is, instead of starting from a representation with $0$ negated terms which is repeteadly populated by negations involving infrequent terms, we start from a logical formula with $T \setminus T_d$ negated terms and we repeteadly omit globally frequent terms[5].

We define now the number of terms that will be omitted in the logical representation of a given document as a function of the size of the document:

$$OT_d = \frac{max\_dl - dl_d}{max\_dl - min\_dl} \cdot MAX\_OT \quad (1)$$

where $dl_d$ is the size of the document $d$, $max\_dl$ ($min\_dl$) is the size of the largest (shortest) document and $MAX\_OT$ is a constant that determines the maximum number

[5] In the future we also plan to articulate an indexing process which skips globally infrequent terms and, hence, these procedures will be revisited.

of terms for which the logical indexing will not make any strong decision and, hence, no literal, either positive or negative, will be expressed[6].

To sum up, every document will be represented as a logical formula in which:

- Terms appearing explicitly in the text of the document, $t \in T_d$, will be positive literals in the representation of the document.
- Terms not mentioned explicitly, $t \in T \setminus T_d$ are ranked in decreasing order of appearances within the whole collection and:
  - Top $OT_d$ terms will be omitted in the representation of $d$.
  - The remaining terms will be negative literals in the logical formula representing $d$.

$$T = \{a, b, c, d, e, f, g, h, i, j, l, m, n, o, p, q, r, s, t, u\}$$

| Document | $T_d$ (explicit terms) |
|---|---|
| $d_1$ | $a, r$ |
| $d_2$ | $a, c, d, e, u, t$ |
| ... | ... |

$$max\_dl = 10$$
$$min\_dl = 2$$
$$MAX\_OT = 10$$

$$OT_{d_1} = \frac{10-2}{10-2} \cdot 10 = 10$$
$$OT_{d_2} = \frac{10-6}{10-2} \cdot 10 = 5$$

| Document | omitted terms | negated terms |
|---|---|---|
| $d_1$ | $u, t, s, q, p, o, n, m, l, j$ | $i, h, g, f, e, d, c, b$ |
| $d_2$ | $s, r, q, p, o$ | $n, m, l, j, i, h, g, f, b$ |
| ... | ... | ... |

| Document | Logical representation |
|---|---|
| $d_1$ | $a \wedge r \wedge \neg i \wedge \neg h \wedge \neg g \wedge \neg f \wedge \neg e \wedge \neg d \wedge \neg c \wedge \neg b$ |
| $d_2$ | $a \wedge c \wedge d \wedge e \wedge u \wedge t \wedge \neg n \wedge \neg m \wedge \neg l \wedge \neg j \wedge \neg i \wedge \neg h \wedge \neg g \wedge \neg f \wedge \neg b$ |
| ... | ... |

**Fig. 1.** Logical indexing process

Figure 1 illustrates an example of this logical indexing process. The vocabulary of 20 terms is supposed to be ordered in increasing order of appearance within the whole collection. The largest document is supposed to have 10 terms whereas the shortest one ($d_1$) mentions just two terms. The constant $MAX\_OT$ is assumed to be equal to 10. Observe that, a *closed-world assumption indexing* would assign 18 and 14 negations to $d_1$ and $d_2$, respectively, whereas the length-dependent indexing assigns 8 negations to

[6] Of course, $MAX\_OT$ should be lower or equal than the smallest value of $|T \setminus T_d|$ computed across all documents. Otherwise, the indexing process could suggest a value of $OT_d$, such that $OT_d > |T \setminus T_d|$. This indexing could only be implemented by considering some explicit terms in $T_d$ as non-informative words that should be omitted the representation of the document. Obviously, this is not the intention pursued here.

| Topics #151-#200 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | cwa indexing | MAX_OT 1000 | MAX_OT 2000 | MAX_OT 3000 | MAX_OT 4000 | MAX_OT 5000 | MAX_OT 10000 | MAX_OT 50000 |
| 0.4 | 0.0719 | 0.1320 | 0.1544 | 0.1475 | 0.1420 | 0.1422 | 0.1136 | 0.0736 |
| | 1533 | 2013 | 2090 | 1849 | 1912 | 1845 | 1639 | 1539 |
| 0.5 | 0.1055 | 0.1470 | **0.1687** | 0.1562 | 0.1537 | 0.1613 | 0.1526 | 0.1075 |
| | 1760 | 2010 | **2048** | 1786 | 1837 | 1950 | 1810 | 1764 |
| 0.6 | **0.1520** | 0.1561 | 0.1513 | 0.1289 | 0.1041 | 0.1290 | 0.1426 | 0.1452 |
| | **1751** | 1864 | 1738 | 1447 | 1298 | 1578 | 1522 | 1748 |

**Table 1.** Training phase - Tuning partiality

the short document $d_1$ and 9 negations to the long document. Note that the final logical representation of a long document is more complete because there will be few omitted terms and, on the contrary, representations of short documents are more partial.

The tuning constant $MAX\_OT$ is an instrument to make explicit control on partiality. If $MAX\_OT = 0$ then the system does not allow partiality in the logical representations and, therefore, all the vocabulary terms have to be mentioned either positive or negative. As $MAX\_OT$ grows logical representations become more partial. Obviously, very low values of $MAX\_OT$ will not permit to establish significant differences between the indexing of short and long documents.

## 4 Experiments

This logical indexing was evaluated against the WSJ subset of the TREC collection in discs 1&2. This collection constains 173252 articles published in the Wall Street Journal between 1987 and 1992.

We took 50 TREC topics for training the $MAX\_OT$ parameter (TREC topics #151 - #200) and a separate set of topics is later used for validating previous findings (TREC topics #101 - #150). For each query, top 1000 documents were used for evaluation.

Documents and topics were preprocessed with a stoplist of 571 common words and remaining terms were stemmed using Porter's algorithm [10]. Logical queries are constructed by simply connecting their stems through logical conjunctions. Queries are long because the subparts Title, Description and Narrative were all considered. Stemmed document terms are directly incorporated as positive literals and some negated terms are included in the conjunctive representation of a document depending on document's length and term's global frequency. In order to check whether or not this new logical indexing improves the top performance obtained by the PLBR model so far, we first ran a number of experiments following a closed-world assumption (i.e. all terms which are not mentioned by the document are incorporated as negated literals). Recall that the PLBR model handles idf information when measuring distances between logical interpretations. This effect is controlled by a parameter $\alpha$. We tried out values for $\alpha$ from 0.9 to 0.1 in steps of 0.1. Since the major benefits were found when $0.4 \leq \alpha \leq 0.6$, we only present performance results for $\alpha = 0.4, 0.5, 0.6$. On the second column of table 1 (cwa indexing) we show performance ratios (non-interpolated average precision & total number of relevant retrieved documents) for the cwa indexing approach on the training set. The best results were found for a value of $\alpha$ equal to 0.6 (in bold).

Columns 3rd to 9th of table 1 depict performance results for the more evolved logical indexing with varied number of omitted terms. Not surprinsingly, for high values of

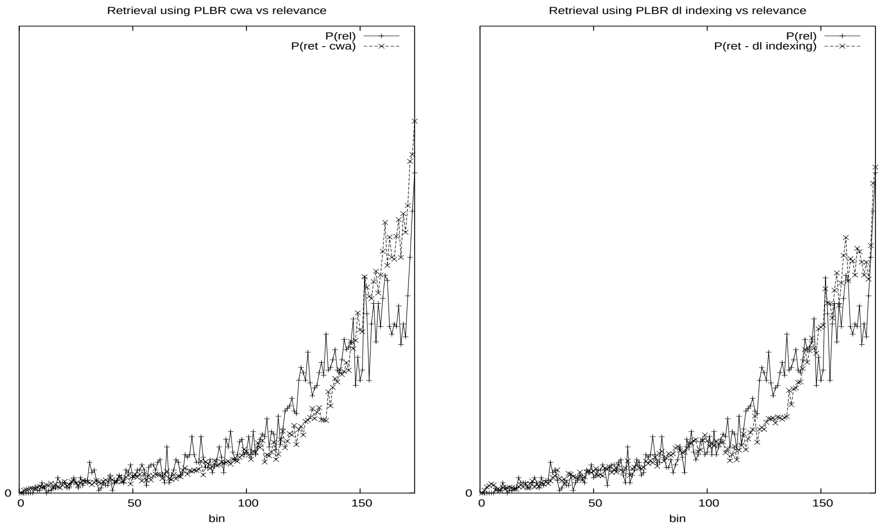
| Topics #101-#150 | | |
| --- | --- | --- |
| Recall | cwa indexing trained $\alpha = 0.6$ | dl indexing not trained $MAX\_OT = 2000$, $\alpha = 0.5$ |
| 0.00 | 0.4576 | 0.5379 |
| 0.10 | 0.2842 | 0.3317 |
| 0.20 | 0.2154 | 0.2639 |
| 0.30 | 0.1788 | 0.2075 |
| 0.40 | 0.1445 | 0.1600 |
| 0.50 | 0.1195 | 0.1240 |
| 0.60 | 0.0923 | 0.0993 |
| 0.70 | 0.0717 | 0.0684 |
| 0.80 | 0.0397 | 0.0373 |
| 0.90 | 0.0188 | 0.0131 |
| 1.00 | 0.0098 | 0.0035 |
| Avg.prec. (non-interpolated) | 0.1319 | 0.1482 |
| % change | | +12.4% |
| Total relevant retrieved | 1828 | 2301 |
| % change | | +25.9% |

**Table 2.** Test phase - Effect of partiality

omitted terms ($\geq 50000$) performance tends to the performance obtained with the basic indexing (first column). This is because the ratio $negated\_terms/omitted\_terms$ is so low that almost every query term is either matched by a document or it was omitted. There are very few negations and, hence, the distinction between those classes of terms is unnoticeable. On the other hand, for relatively low values of $MAX\_OT$ (between 1000 and 5000) performance tends to improve with respect to cwa indexing. The best training run is obtained when $MAX\_OT = 2000, \alpha = 0.5$ ($0.1687$ vs $0.152$, $11\%$ improvement in non-interpolated average precision and 2048 vs 1751, 17% more relevant documents retrieved).

In order to confront previous findings, we ran additional experiments with the test set of topics. We fixed a value of 2000 omitted terms and $\alpha = 0.5$ for the new indexing approach. Although this is the test phase, we trained again the parameter $\alpha$ for the basic indexing policy (cwa) and we show here the best results ($\alpha = 0.6$). This is to assure that the new document length indexing without training can improve the best results attainable with the basic cwa indexing. The results are depicted in table 2. Major benefits are found when partial representations are handled. It seems clear that the consideration of document length to omit up to 2000 terms improves significantly the retrieval performance of the logical model.

This experimentation suggests to omit a relatively low number of omitted terms with respect to the total vocabulary size. This means that the shortest document will be able to have 2000 omitted terms within its logical representation. These 2000 terms will be those more globally used that are not present in that small document. It is well known [12] that the large majority of the words occurring in a corpus have very low document frequency. This means that most terms are used just once in the whole collection and, hence, it is also unlikely that any query makes use of them. That is, only a small fragment of the vocabulary (the most frequent ones) makes a significant impact on retrieval performance. Indeed, in the WSJ collection that we indexed, 76839 terms out of 163656 (which is the vocabulary size after preprocessing) are only mentioned in a single document. This explains why the major differents in performance are found for small values of $MAX\_OT$.




(a) relevance vs PLBR cwa (b) relevance vs PLBR dl indexing

**Fig. 2.** Probability of relevance vs probability of retrieval

## 5 Analysis

In this section we provide an additional analysis of the logical indexing keeping track of its behaviour against document length. We will follow the methodology designed by Singhal, Buckley and Mitra [13] to analyze the likelihood of relevance/retrieval for documents of all lengths and plot these likelihoods against the document length to compare the relevance pattern and the retrieval pattern.

First, the document collection is ordered by document's length and documents are divided into equal-sizes chunks, which are called bins. For our case, the 173252 WSJ documents were divided into 173 bins containing one thousand documents each and an additional bin contained the 252 largest documents. For the test topics (#101-#150) we then took the 4556 (query, relevant document) pairs and counted how many pairs had their document from the ith bin. These values allow to plot a relevance pattern against document length. Specifically, the conditional probability $P(D \in ith\; bin | D\; is\; relevant)$ can be computed as the ratio of the number of pairs that have the document from the ith bin and the total number of pairs.

A given retrieval strategy will present a good behaviour against document's length provided that its probability of retrieval for the documents of a given length is very close to the probability of finding a relevant document of that length. Therefore, once we have a relevance pattern, we can compute the retrieval pattern and compare them graphically. We will compute the retrieval pattern for both the cwa PLBR run and the PLBR run with document length-dependent indexing. Comparing them with the relevance pattern we will be able to validate the adequacy of our document's length-dependent indexing and, possibly, identify further avenues of research.

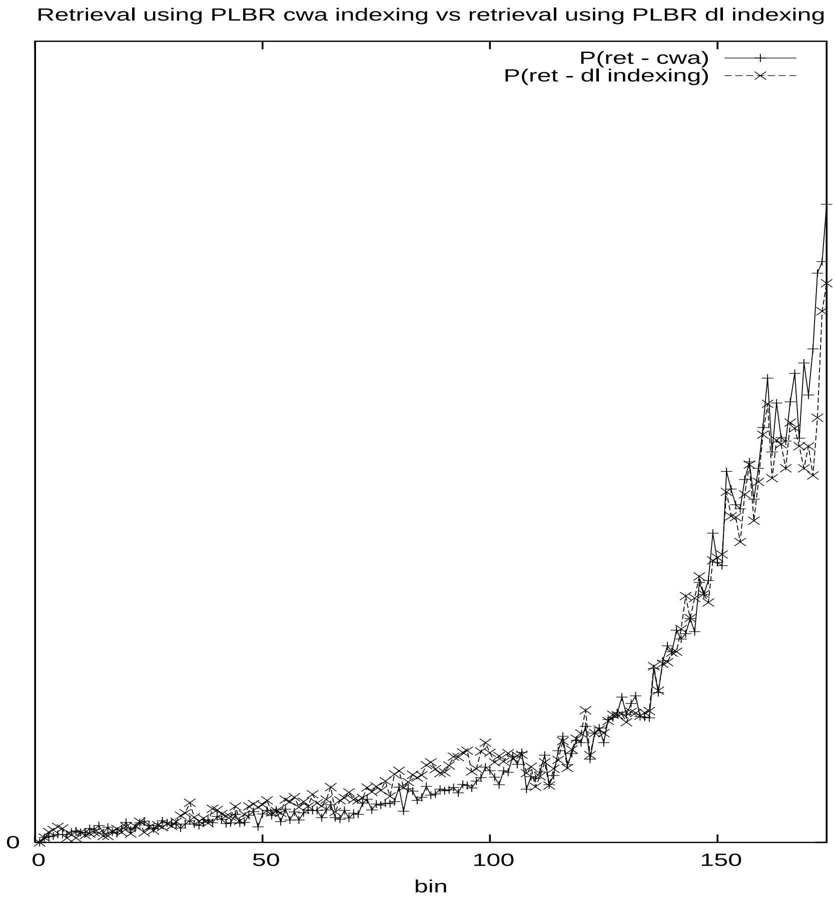



**Fig. 3.** cwa indexing vs dl indexing

The retrieval pattern's computation is also very simple. For each query the top one thousand documents retrieved are selected (for our case, 50.000 (query, retrieved documents) pairs) and, for each bin, we can directly obtain $P(D \in ith\, bin | D\, is\, retrieved)$.

Figure 2 shows the probability of relevance and the probability of retrieval of the cwa PLBR run plotted against the bin number (2(a)). The probability of relevance and the probability of retrieval applying the document length-dependent logical indexing are plotted in fig. 2(b). Recall that bin #1 contains the smallest documents and bin #174 containst the largest documents. Following that figure, there is no clear evidence about the distinction between both approaches. In figure 3 we plot cwa indexing and dl indexing against document length. Although the curves are very similar, some trends can be identified. For bins #1 to #100 the dl indexing approach retrieves documents with higher probability than the cwa approach. On the other hand, very long documents (last 20 bins) are retrieved with higher probability if the cwa strategy is applied. This demonstrates that the dl indexing procedure does its job because it tends to favour short documents w.r.t. long ones. Nevertheless, this analysis also suggests new ways to improve the document length logical indexing. The most obvious is that very long documents do still present a probability of being retrieved which is much greater than the probability of relevance (see fig. 2(b), last 20 bins). This suggests that the formula that computes the number of omitted terms (equation 1, section 3.2) should be adapted accordingly. As a consequence, subsequent research effort will be directed to the fine tuning of the document length-dependent indexing.

## 6 Conclusions and future work

In this work we have proposed a novel logical indexing technique which yields a natural way to handle terms not explicitly mentioned by documents. The new indexing approach is assisted by popular IR notions such as document length normalization and global term distribution. The combination of those classical notions and the expressiveness of the logical apparatus leads to a precise modeling of the document's contents. The evaluation conducted confirms empirically the advantages of the approach taken.

Future work will be focused in a number of lines. First, as argued in the previous section, document length contribution should be optimized. Second, more evolved techniques to negate terms will also be investigated. In this respect, the application of term similarity information is especially encouraging for avoiding negated terms whose semantics is close to some of the terms which appear explicitly in the text of a document.

Our present document length strategy captures verbosity by means of document length. Although it is sensible to think that there is a correlation between document length and verbosity, it is also very interesting to study new methods to identify verbose/scope documents and tune the model accordingly.


## Acknowledgements

This work was supported by projects TIC2002-00947 (from "Ministerio de Ciencia y Tecnología") and PGIDT03PXIC10501PN (from "Xunta de Galicia"). The first author is supported in part by "Ministerio de Ciencia y Tecnología" and in part by FEDER funds through the "Ramón y Cajal" program.


## References


1. F. Crestani, M. Lalmas, and C. J. van Rijsbergen (editors). *Information Retrieval, Uncertainty and Logics: advanced models for the representation and retrieval of information.* Kluwer Academic, Norwell, MA., 1998.
2. W. B. Croft and J. Lafferty. *Language Modeling for Information Retrieval.* Kluwer Academic, 2003.
3. M. Dalal. Investigations into a theory of knowledge base revision:preliminary report. In *Proceedings of the 7th National Conference on Artificial Intelligence (AAAI'88)*, pages 475–479, Saint Paul, USA, 1988.

4. D. Losada and A. Barreiro. Embedding term similarity and inverse document frequency into a logical model of information retrieval. *Journal of the American Society for Information Science and Technology, JASIST*, 54(4):285–301, February 2003.
5. D. Losada and A. Barreiro. Propositional logic representations for documents and queries: a large-scale evaluation. In F. Sebastiani, editor, *Proc. 25th European Conference on Information Retrieval Research, ECIR'2003*, pages 219–234, Pisa, Italy, April 2003. Springer Verlag, LNCS 2663.
6. D. E. Losada and A. Barreiro. Using a belief revision operator for document ranking in extended boolean models. In *Proc. SIGIR-99, the 22nd ACM Conference on Research and Development in Information Retrieval*, pages 66–73, Berkeley, USA, August 1999.
7. D. E. Losada and A. Barreiro. Efficient algorithms for ranking documents represented as dnf formulas. In *Proc. SIGIR-2000 Workshop on Mathematical and Formal Methods in Information Retrieval*, pages 16–24, Athens, Greece, July 2000.
8. D. E. Losada and A. Barreiro. A logical model for information retrieval based on propositional logic and belief revision. *The Computer Journal*, 44(5):410–424, 2001.
9. J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. 21st ACM Conference on Research and Development in Information Retrieval, SIGIR'98*, pages 275–281, Melbourne, Australia, 1998.
10. M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
11. S. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. SIGIR-94, the 17th ACM Conference on Research and Development in Information Retrieval*, pages 232–241, Dublin, Ireland, July 1994.
12. G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.
13. A Singhal, C. Buckley, and M Mitra. Pivoted document length normalization. In *Proc. SIGIR-96, the 19th ACM Conference on Research and Development in Information Retrieval*, pages 21–29, Zurich, Switzerland, July 1996.
14. C. J. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2):106–119, 1977.
15. C.J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29:481–485, 1986.
16. H. Zaragoza, D. Hiemstra, and M. Tipping. Bayesian extension to the language model for ad hoc information retrieval. In *Proc. 26th ACM Conference on Research and Development in Information Retrieval, SIGIR'03*, pages 4–9, Toronto, Canada, 2003.