Revisiting Logical Imaging for Information Retrieval

Guido Zuccon, Leif Azzopardi, Cornelis J. van Rijsbergen
Dept. of Comp. Sci., University of Glasgow
Glasgow, United Kingdom
{guido, leif, keith}@dcs.gla.ac.uk

ABSTRACT

Retrieval with Logical Imaging is derived from belief revision and provides a novel mechanism for estimating the relevance of a document through logical implication (i.e. $P(\mathbf{q} \rightarrow \mathbf{d})$). In this poster, we perform the first comprehensive evaluation of Logical Imaging (LI) in Information Retrieval (IR) across several TREC test Collections. When compared against standard baseline models, we show that LI fails to improve performance. This failure can be attributed to a nuance within the model that means non-relevant documents are promoted in the ranking, while relevant documents are demoted. This is an important contribution because it not only contextualizes the effectiveness of LI, but *crucially explains why it fails*. By addressing this nuance, future LI models could be significantly improved.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval - *Retrieval Models*
General Terms: Theory, Experimentation
Keywords: Logical Imaging, Probability kinematics

1. INTRODUCTION

Logical Imaging (LI) is a technique for belief revision which has been employed in a retrieval model introduced by Crestani *et al* [3]. They proposed to derive the probability of relevance $\mathbf{R}$ of a document $\mathbf{d}$ given a query $\mathbf{q}$, namely $P(\mathbf{R}|\mathbf{d}, \mathbf{q})$, by computing the probability of a conditionalization, in particular $P(\mathbf{d} \rightarrow \mathbf{q})$ (Imaging on the document) or $P(\mathbf{q} \rightarrow \mathbf{d})$ (Imaging on the query). While the original proposal of applying LI in IR was put forward by van Rijsbergen [4] in 1986, it took several years before Amati *et al* [1] and Crestani *et al* [3] provided the first retrieval methods using LI. The first implemented the latter conditionalization, $P(\mathbf{q} \rightarrow \mathbf{d})$, while the second used the conditionalization $P(\mathbf{d} \rightarrow \mathbf{q})$.

In this poster, we focus on the latter method which was studied in somewhat more, but still limited, depth. Briefly, this technique assumes that a probability[1] is associated with each term in the collection. Successively, this LI method scores a document $\mathbf{d}$ summing not only the probabilities of terms appearing both in a query and in the document $\mathbf{d}$, but also the probabilities of terms which are considered similar to the query terms belonging to $\mathbf{d}$[2]. The intuition behind the model is that the probability mass of similar terms will be moved into the document and increase the document's relevance if more similar to the query terms. In [3] a number of experiments were conducted on small test collections (i.e. CACM and CRAN) in an attempt to determine whether this intuition would improve retrieval performance. While these experiments reported improvements over an IDF baseline, the model was never compared to models such as TF.IDF or BM25. Nor was it ever tested on large scale test collections. Furthermore, no working implementation exists.

In this work, we revisit LI and perform a comprehensive evaluation of the model on several TREC test collections. The remainder of this poster is structured as follows: in the next section we provide an overview of LI applied to document Imaging [3]. Then in Section 3, we outline the set of experiments undertaken as part of this study along with the results of the experiments. The poster concludes in Section 4, stating the main contribution of this work and the possible avenues for future investigation.

2. LOGICAL IMAGING IN IR

A probability distribution P on the set T of terms in a collection D is initially defined such that the sum of all terms probabilities is one. Each document $\mathbf{d}$ is represented using terms belonging to T. A document can either be true or not true in the context of a term, i.e. the document either contains or not the term. In order to evaluate the probability of the conditionalization, namely $P(\mathbf{d} \rightarrow \mathbf{q})$, LI on $\mathbf{d}$ is applied, leading to the following computation:

$$P(\mathbf{d} \rightarrow \mathbf{q}) = P_d(\mathbf{q}) = \sum_{t \in T} P(t) t_d(\mathbf{q})$$

where $t_d(\mathbf{q})$ is the truth function which returns 1 if and only if $\mathbf{q}$ is true at t_d, 0 otherwise, and t_d is the most similar term to t for which $\mathbf{d}$ is true. The term similarity is usually computed by means of the co–occurrences of the two terms. In particular, the Expected Mutual Information Measure $(EMIM)$[3] has been used in [3] and is adopted in this work as well.

[1]In [3], this probability is approximated by the IDF weight of the term, being the set of IDF weights of term in the collection monotonical to a probability distribution.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

[2]Note, this makes the method computationally expensive, i.e. every term outside the document must be every time compared to the terms inside the document [2].

[3]The similarity between term t_i and term t_j is defined as $EMIM(t_i, t_j) = \sum_{t_i, t_j} P(t_i, t_j) \log \frac{P(t_i, t_j)}{P(t_i)P(t_j)}$, where $P(t_i, t_j)$ is the probability the two terms co–occurs together in a windows of text, usually set to have length 10.

3. METHOD AND RESULTS

The following experimental comparison was performed on three TREC Collections: Associated Press (AP8889) and Wall Street Journal (WSJ8792) using TREC 1, 2, 3 Topics, and the Los Angeles Times (LA8990) using the TREC 6, 7, 9 Topics. Each collection was indexed using Lemur, where they were stemmed and stopped. The LI Model was also implemented in Lemur. Given the previous implementation problems regarding efficiency [2], this implementation re-ranks the IDF baseline, since the IDF baseline is essentially LI without Imaging. In our experiments the top 1,000 documents were re-ranked. For each collection, we compared LI, against IDF, TF.IDF and BM25, where significance testing was performed using the t-test ($p < 0.05$). Table 1 presents a summary of results. While LI provides some improvement over IDF, it is significantly and substantially outperformed by TF.IDF and BM25 across MAP, p@10, p@20, and bpref for all collections. Clearly, LI is inferior to standard baseline models.

Collection	Model	MAP	P10	P20	bpref
WSJ8792	**LI**	11.96	23.40	20.40	19.69
	IDF	8.41†	21.66†	20.38	13.06†
TREC 1,2,3	**TF.IDF**	25.16*	42.62*	37.79*	30.28*
	BM25	26.06*	44.30*	39.16*	31.60*
AP8889	**LI**	6.63	19.26	17.01	19.57
	IDF	5.60	14.55†	14.66†	14.55†
TREC 1,2,3	**TF.IDF**	17.24*	28.12*	26.68*	33.03*
	BM25	18.24*	28.59*	27.68*	33.96*
LA8990	**LI**	11.98	16.13	12.70	14.72
	IDF	10.40†	15.00	13.17	12.33†
TREC 6,7,8	**TF.IDF**	22.06*	25.00*	19.20*	21.84*
	BM25	21.57*	25.93*	20.13*	20.74*

Table 1: Values in percentage of MAP, P10, P20 and bpref for LI, IDF, TF.IDF and BM25 using WSJ8792, AP8889 and LA8990. Significance is denoted by * which indicates method X is better than LI, while † indicates LI is better than method X.

4. DISCUSSION AND CONCLUSION

Despite the theoretical soundness, the current formulation of LI is not effective in retrieval tasks. After examining the scoring function, we identified a nuance within the model that contributes to its ineffectiveness. The following example highlights the problem. Consider documents $\mathbf{d_1} = \{pet, cat, dog, bird, shop\}$ and $\mathbf{d_2} = \{fish, chip, shop\}$, where we have indicated in brackets the terms present in each document. Given a IDF or TF.IDF retrieval system, a user submitting the query $\mathbf{q} = \{pet, shop\}$ will receive a list of documents where $\mathbf{d_1}$ is ranked higher than $\mathbf{d_2}$, since $P(\mathbf{R}|\mathbf{d_1}, \mathbf{q}) > P(\mathbf{R}|\mathbf{d_2}, \mathbf{q})$ (see Eq. 1 and Eq. 2). However, LI would revise the initial probability of Relevance given each document and $\mathbf{q}$, namely Eq. 1 and Eq. 2. The initial probabilities are revised in accordance with a probability kinematics policy that transfers probabilities from terms absent from the document to terms present in it. The selection of terms subjected to such kinematics is performed employing EMIM. In this example, for document $\mathbf{d_1}$ there is a transfer of probabilities from terms *fish*, *chip* to *shop*, since *fish* and *chip* have a higher similarity to *shop* than to any other term in $\mathbf{d_1}$. Thus, the revised probability of *shop* in $\mathbf{d_1}$ is $P'_{d_1}(shop) = P(shop) + P(fish)$. Consequently, LI retrieves document $\mathbf{d_1}$ in response to $\mathbf{q}$ with a score proportional to Eq. 3. Similarly, in document $\mathbf{d_2}$ term *shop* is the attractor of probabilities transfers having origins from terms *pet, cat, dog, bird*, leading to $P'_{d_2}(shop) = P(shop) + P(pet) + P(cat) + P(dog) + P(bird)$. Thus, $\mathbf{d_2}$ would be retrieved by LI with a score proportional to Eq. 4.

Before Imaging:

$$P(\mathbf{R}|\mathbf{d_1}, \mathbf{q}) \approx P(pet) + P(shop) \quad (1)$$

$$P(\mathbf{R}|\mathbf{d_2}, \mathbf{q}) \approx P(shop) \quad (2)$$

After Imaging:

$$P(\mathbf{d_1} \rightarrow \mathbf{q}) \approx P(shop) + P(pet) + P(fish) + P(chip) \quad (3)$$

$$P(\mathbf{d_2} \rightarrow \mathbf{q}) \approx P(shop) + P(pet) + P(cat) + P(dog) + P(bird) \quad (4)$$

Figure 1: Equations representing the probabilities associated to documents $\mathbf{d_1}$ and $\mathbf{d_2}$ of the example presented in section 4 before and after the Imaging process.

Comparing the scores associated with the two documents, LI ranks $\mathbf{d_2}$ higher than $\mathbf{d_1}$, since Eq. 4 > Eq. 3.[4] Clearly, this is not desirable as non relevant information is promoted above relevant.

Another problem stems from the fact that a document containing related terms does not benefit from the transfers, as only external terms are transfered into the document. This tends mean that shorter documents tend to be favored over longer. Since length normalization is important for retrieval functions it is likely that this problem also contributes to poorer retrieval performance.

In conclusion, in this poster we have revisited LI performing the first, thorough and comprehensive evaluation of the model proposed in [3] and provide a working implementation[5] The major findings of this work are: (1) LI does not significantly improve the overall performance (map and bref), (2) it is significantly worse than standard retrieval models like TF.IDF and BM25, and (3) the LI Model is affected by a nuance in the scoring function which demotes relevant documents. Nonetheless, LI does provide some improvement at early precision which suggests that it may have some potential if the problems in the ranking function can be addressed. Further work will explore these differences, along with re-considering how LI can be applied within a state of the art model, such as the Language Model, which appears to be naturally suited to the belief revision process.

5. ACKNOWLEDGMENTS

The authors would like to thank W. Vanderbauwhede, and IRF[6]. This work is partially funded by EPSRC EP/F014384/.

6. REFERENCES

[1] G. Amati and S. Kerpedjiev. An Information Retrieval Logic Model: Implementation and Experiments. Technical Report Rel5B04892, FUB, Italy, 1992.

[2] F. Crestani, I. Ruthven, M. Sanderson, and C. J. van Rijsbergen. The Troubles with Using a Logical Model of IR on a Large Collection of Documents. In *Proc. TREC4*, pages 509–526, 1995.

[3] F. Crestani and C. J. van Rijsbergen. Probability Kinematics in Information Retrieval. *Proc. ACM SIGIR*, pages 291–299, 1995.

[4] C. J. van Rijsbergen. A new Theoretical Framework for Information Retrieval. *Proc. ACM SIGIR*, pages 194–200, 1986.

[4]Note that since we assume $P(t) = idf(t)$, then $P(fish) = P(pet) = P(cat) = P(dog) = P(bird)$.

[5]Available from the first author by request.

[6]http://www.ir-facility.org/