

A possibilistic logic based information retrieval model with various term-weighting approaches

Janusz Kacprzyk¹, Katarzyna Nowacka², and Sławomir Zadrozny¹

¹ Systems Research Institute PAS, ul. Newelska 6, 01-447 Warsaw, Poland,

² Doctoral Studies (SRI PAS), ul. Newelska 6, 01-447 Warsaw, Poland

Abstract. A new, possibilistic logic based information retrieval model is presented. Its main feature is an explicit representation of both the vagueness and the uncertainty typical for the textual information representation and processing. The weights of index terms in documents and queries are directly interpreted as quantifying this vagueness and uncertainty. The classical approaches to the term-weighting are tested on a standard data set in order to validate their appropriateness for expressing vagueness and uncertainty³.

1 Introduction

A model of information retrieval consists in establishing a representation of both documents and query(ies) as well as a mechanism for their matching. For these purposes logical models [1, 2] employ logical formulae for the representation and some notions related to the logical entailment as the matching mechanism. Many approaches of this type are known in the literature that use various types of logic as their theoretical foundations (for the reviews cf., e.g., [3, 4]).

In this paper we propose a possibilistic logic based approach that makes possible an explicit representation of the vagueness and uncertainty pervading the information processing. In particular, using an extended version of the possibilistic logic we directly reflect in the representation of documents and queries both the vagueness of the relevance concept as well as uncertainty related to its assessment.

2 Basics of the possibilistic logic and its extension

The possibilistic logic has been introduced by Dubois, Lang and Prade [5, 6]. Here we recall only its basic concepts and limit our discussion to its propositional version. Moreover we are interested mainly in its semantics.

The starting point is the classical propositional logic which is extended by the introduction of weighted formulae. The motivation for the introduction of weights is to make it possible to directly express the uncertainty as to the validity of a formula: the higher the weight the more certain we are as to the validity of

³ Research supported by the KBN Grant 3 T11C 052 27

the formula. This (un)certainty is modeled in the framework of the possibility theory [7, 8] and might be conveniently explained referring to the classical logical notions of the *interpretation* and *model*.

Let us assume an alphabet of our language, i.e., a set \mathcal{A} of the propositional variables. Then, an interpretation $\omega \in \Omega$ is a function:

$$\omega : \mathcal{A} \rightarrow \{0, 1\}$$

assigning the truth values 1 (“true”) or 0 (“false”) to all propositional variables of the alphabet \mathcal{A} . In the classical logic, for a given formula p we distinguish those interpretations Ω^p that make formula p true and we call $\omega \in \Omega^p$ *models* of p . Thus assuming p true makes the interpretations belonging to Ω^p possible and those belonging to $\Omega \setminus \Omega^p$ impossible. Using the language of possibility theory we will say that p induces on the set Ω the following possibility distribution:

$$\pi_p(\omega) = \begin{cases} 1 & \forall \omega \in \Omega^p \\ 0 & \forall \omega \notin \Omega^p \end{cases} \quad (1)$$

In the propositional possibilistic logic the formulae take the following form:

$$(p, \alpha) \quad (2)$$

where $\alpha \in [0, 1]$ expresses the lower bound on the belief in truth of p while p is still assumed to be either “true (1)” or “false (0)”. Such a weighted formula (p, α) induces the following possibility distribution over the set of interpretations Ω (cf. (1)):

$$\pi_p(\omega) = \begin{cases} 1 & \text{if } \omega \in \Omega^p \\ 1 - \alpha & \text{if } \omega \notin \Omega^p \end{cases} \quad (3)$$

Having (p, α) assumed the belief in a truth of any formula q is calculated as the pair of the possibility and necessity measures (induced by the distribution π_p (3)) of the set of q ’s models, i.e., as $(\Pi(\Omega^q), \mathcal{N}(\Omega^q))$. Moreover, if we consider a set of weighted formulae $P = \{(p_i, \alpha_i)\}$ then jointly they induce the following possibility distribution on Ω :

$$\pi_P(\omega) = \min_i \pi_{p_i}(\omega) \quad (4)$$

We will use the weighted formulae (2) to express the uncertainty of the relevance of an index term for a document or query. As we assume the relevance to be a vague concept we will use an extension of the possibilistic logic for the many valued case as proposed by Lehmke [9]. In this approach the formulae are still weighted formulae, but the weights are interpreted differently and referred to as *labels*. A label l is interpreted as a fuzzy set of truth values and (p, l) induces a possibility distribution π_p on the set of interpretations $\Omega = \{\omega : \mathcal{A} \rightarrow [0, 1]\}$ such that:

$$\pi_p(\omega) = \mu_l(\omega(p)) \quad (5)$$

For a set of formulae $\{(p_i, l_i)\}$ (*the possibilistic knowledge base*) a possibility distribution induced by all them jointly, is calculated as previously using (4).

Some special labels of interest here are as follows [9]. The label l^T referred to as “TRUE” has the membership function:

$$\mu_{l^T}(x) = x \quad (6)$$

Thus assuming (p, l^T) means that more possible are interpretations giving p a higher truth value. A slightly modified version of l^T referred to as “TRUE with doubt δ ” and denoted as l_δ^T is defined by this membership function:

$$\mu_{l_\delta^T}(x) = \begin{cases} \delta & \forall x \leq \delta \\ x & \forall x > \delta \end{cases} \quad (7)$$

Assuming (p, l_δ^T) the belief in p is somehow limited, thus the possibility of interpretations assigning to p the truth value lower than δ is equal δ instead of reducing towards 0 as it is the case of l^T .

3 The model

We use the following notation: $D = \{d_i\}_{i \in [1, N]}$ is the set of documents and $T = \{t_j\}_{j \in [1, M]}$ the set of index terms. Following the logical approach there is a propositional variable p_j corresponding to each index term t_j . Then, each document/query is represented as a set of weighted propositions $d = \{(p_j, l)\}$ which is interpreted as the possibilistic knowledge base (cf. previous section).

Whatever means, either manual or automatic, are used to establish the relevance relation between the index terms and the documents/queries one cannot be completely sure as to the results - this is the source of uncertainty that we want to model. On the other hand the very notion of relevance is gradual rather than binary. Representing a document/query with a set of weighted formulae (p_j, l) , where p_j corresponds to an index term t_j , makes it possible to model both uncertain and vague nature of the relevance between index terms and documents/queries.

Documents Firstly, each document is represented as in the vector space model using selected term-weighting approach [10]). Thus, for each document $d \in D$ and each index term $t_j \in T$ a weight $d(t_j)$ is computed. Then the weights are normalized so as to obtain the maximum weight equal 1. Finally, the document d is represented as a set of weighted formulae (8):

$$\{(p_j, l(d(t_j)))\}_{j=1, \dots, M} \quad (8)$$

where p_j is a propositional variable corresponding to the index term t_j and $l(u)$ is a label with the following membership function (parametrized with u):

$$\mu_{l(u)}(x) = 1 - |x - u| \quad (9)$$

Thus, recalling the semantics of weighted formulae, the induced possibility distribution π_d^j on the interpretations $\omega \in \Omega$ “favors” those that are assigning p_j

the truth value close to $d(t_j)$. The larger the difference $\omega(p_j) - d(t_j)$ the less possible the interpretation ω . The overall possibility distribution induced by the document d represented with the possibilistic knowledge base (8) is computed using (4).

The motivation for (8)-(9) is such that it is assumed that the weight $d(t_j)$ expresses the relevance of t_j for the representation of d . However due to the uncertainty of the indexing process other levels of relevance are also possible to a degree quantified by (9).

Queries For a query q and each index term $t_j \in T$ the normalized weight $q(t_j)$ is computed as in case of documents, although a different term-weighting approach might be applied. Finally the query q is represented as the following possibilistic knowledge base:

$$\{(p_j, l_{1-q(t_j)}^T)\}_{j=1, \dots, M} \quad (10)$$

where p_j is a propositional variable corresponding to index term t_j and $l_{1-q(t_j)}^T$ is a label of the l_δ^T type (cf. (7)) with $\delta = 1 - q(t_j)$.

The possibility distribution on $\omega \in \Omega$ induced by the possibilistic knowledge base (10) representing the query q favors those ω that assign to p_j as high truth degree as possible. However, those ω that assign low truth values are still possible to some degree which is the higher the lower $q(t_j)$ is.

The motivation for (10) is such that it is assumed that the index terms used in the query are treated by the user as fully relevant (to the degree 1). It may be argued that in case of usually short queries (typical for, e.g., search engines) there is no need to use partially relevant terms. However, the user might be uncertain if they are really relevant (at all!) and thus $q(t_j)$ is interpreted as an expression of this (un)certainty: when $q(t_j) = 1$ the user is completely sure and l_δ^T in (10) becomes l^T making term t_j highly desired, while small $q(t_j)$, close to 0, expresses high uncertainty as to the relevance of t_j and thus limiting its influence on the resulting possibility distribution over Ω .

Matching degree In order to compute the matching degree between a document and a query we assume the following logical setting. A document is represented as a possibilistic knowledge base (8) and induces the possibility distribution π_d over the set of interpretations Ω defined with (9) and (4). Similarly a query is represented as a possibilistic knowledge base (10) and induces a corresponding possibility distribution π_q that in turn is interpreted as defining a fuzzy set Ω^q of the models of the query q such that $\mu_{\Omega^q}(\omega) = \pi_q(\omega)$. Then, the matching degree of a document d against a query q is computed as the pair of possibility and necessity measures (induced by possibility distribution π_d) of the fuzzy set Ω^q . Formally, that might be expressed as follows:

$$\Pi_d^j(\Omega^q) = \begin{cases} 1 - q(t_j) & \text{for } q(t_j) \leq \frac{1-d(t_j)}{2} \\ \frac{1+d(t_j)}{2} & \text{for } q(t_j) > \frac{1-d(t_j)}{2} \end{cases} \quad (11)$$

$$\mathcal{N}_d^j(\Omega^q) = \begin{cases} 1 - q(t_j) & \text{for } q(t_j) \leq 1 - \frac{d(t_j)}{2} \\ \frac{d(t_j)}{2} & \text{for } q(t_j) > 1 - \frac{d(t_j)}{2} \end{cases} \quad (12)$$

where $d(t_j)$ and $q(t_j)$ are weights of index term t_j in a document d and a query q , respectively; Π_d^j and \mathcal{N}_d^j are possibility and necessity measures induced by π_d^j related to an index term t_j .

The overall matching is expressed as a pair:

$$(\Pi_d(q), \mathcal{N}_d(q)) \quad (13)$$

such that $\Pi_d(q) = \min_j \Pi_d^j(\Omega^q)$ and $\mathcal{N}_d(q) = \min_j \mathcal{N}_d^j(\Omega^q)$.

Having such a pair of numbers it has to be decided how to order the documents in response to the query. In the experiments reported in the next section we assume the lexicographic order on the pairs (13).

4 Experiments

Ideally the labels l in the representations of both documents (8) and queries (10) should be directly determined by a human user, possibly supported by a suitable user interface. However in case of documents it is a rather infeasible solution while in case of queries might require an advanced user interface. Thus in our preliminary experiments with the proposed possibilistic information retrieval model we use the weights computed using some well known term-weighting approaches [10] as a basis for the labels appearing in (9) and (10).

In our experiments we are using the Cranfield test collection (cf. [11] for a description) comprising of 1398 documents and 225 queries. We evaluate the effectiveness of the retrieval using the R-precision measure for each query, i.e., computing the precision of the results at the k -th position of the results list, where k is equal to the number of relevant documents for given query (the list of relevant documents for each query is given as a part of the Cranfield collection). Finally, the R-precision is averaged over all queries.

Table 1 lists the results of our experiments for a few selected combinations of term-weighting approaches for documents/queries. The best results has been obtained for the most popular $\text{tf} \times \text{IDF}$ approach (coded $\text{tfx} \times \text{tfx}$). For a comparison, using the same term-weighting approach and the classical vector space model matching via the cosine measure we have obtained slightly better results (0.2404).

5 Concluding remarks

We have proposed a new possibilistic model of information retrieval. Its main feature is an explicit representation of the vagueness and uncertainty of the relevance. The results of some preliminary computational experiments using various term-weighting approaches are reported. The results are not conclusive and definitely a further research is needed to decide how to proceed with the interpretation of weights as indicators of the uncertainty and vagueness of the relevance relation between index terms and documents/queries.

Documents	Queries	R-Precision	
		(P,N)	(N,P)
tfx	tfx	0.1782	0.1904
tfc	nfx	0.1097	0.1106
nxx	bpc	0.0581	0.0542

Table 1. R-precision for the Cranfield test collection and various term-weighting approaches. The first and second columns indicate term-weighting approaches used for docs/queries representation (coding of approaches as in [10]). The third column shows R-precision when resulting documents are ordered first according to the possibility measure and then to the necessity (cf.(13)). Results in column four are obtained for ordering first according to necessity and then possibility.

References

1. van Rijsbergen C. J.: A new theoretical framework for information retrieval. In Rabitti F., ed.: Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy (1986) 194–200
2. van Rijsbergen C.J.: A non-classical logic for information retrieval. The Computer Journal **29**(6) (1986) 481–485
3. Lalmas M.: Logical models in information retrieval: Introduction and overview. Information Processing & Management **34**(1) (1998) 19–33
4. Sebastiani F.: A note on logic and information retrieval. In: MIRO'95 Proc. of the Final Workshop on Multimedia Information Retrieval, Glasgow, Scotland, Springer (1995)
5. Dubois D., Lang J., Prade H.: Possibilistic logic. In Gabbay D.M. et al., ed.: Handbook of Logic in Artificial Intelligence and Logic Programming. Volume 3. Oxford University Press, Oxford, UK (1994) 439–513
6. Dubois D. and Prade H.: Possibilistic logic: a retrospective and prospective view. Fuzzy Sets and Systems **144** (2004) 3–23
7. Zadeh L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems **1** (1978) 3–28
8. Dubois D., Prade H.: Possibility Theory. Series D: System Theory, Knowledge Engineering and Problem Solving. Plenum Press, New York (1988)
9. Lehmke S.: Degrees of truth and degrees of validity. In Novak V., Perfilieva I., eds.: Discovering the World with Fuzzy Logic. Physica-Verlag, Heidelberg New York (2000) 192–236
10. Salton G. and Buckley C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24** (1988) 513–523
11. Sparck Jones K., Bates R. G.: Research on automatic indexing 1974- 1976 (2 volumes). Technical report, Computer Laboratory. University of Cambridge (1977)