Query Propagation in Possibilistic Information Retrieval Networks

Asma H. Brini
Université Paul Sabatier
brini@irit.fr

Luis M. de Campos
Universidad de Granada
lci@decsai.ugr.es

Didier Dubois
Université Paul Sabatier
dubois@irit.fr

Mohand Boughanem
Université Paul Sabatier
boughane@irit.fr

Abstract

This paper proposes a new Information Retrieval Model based on possibility and necessity measures. This model encodes relationship dependencies existing between terms, documents and query by possibilistic networks. The user's query triggers a propagation process to retrieve necessarily or at least possibly relevant documents.
Keywords: Information Retrieval, Possibilistic Networks, Bayesian Networks, Relevance.

1 Introduction

Information Retrieval process consists in selecting among a large collection those documents that are relevant to a user's query. The relevance of a document to a query is usually interpreted by most of IR models, vector space [9], probabilistic [7][8][12], inference network [14], as a score computed by summing the inner products of term weights in the documents and query representations. The term weighting scheme is indeed the fundamental component of most IR models. The weights are usually based on a combination of measures like term importance in the document (tf), term discrimination in the whole collection of documents (idf) and document length (l_d). These measures, by lack of deeper information, result from frequentist evaluations based on counting of terms. The second important component is the matching strategies that can be used to evaluate document and query representations. This paper proposes an extension of the approach based on possibilistic networks [3]. The possibility approach allow to separate reasons for rejecting a document as irrelevant from reasons to select it by means of two evaluations: possibility and necessity. The present extension results from difficulties to find an efficient way of querying the system. It is too restrictive (and demanding) to aggregate query terms by an AND operator when the only information we have is a set of terms. Thus, the idea is to aggregate query terms by conjunction or disjunction operators according to different aggregation methods when no information is given about the logical description of query. To provide for such a flexibility, a query node is required in the model architecture. We present a general possibilistic approach for IR. A comparative example of query evaluation with existing known models and the proposed model is provided.

2 A possibilistic IR model

Our approach is based on possibilistic directed acyclic networks [1][2], where relations between documents, query and term nodes are quantified by possibility and necessity measures.

2.1 Model architecture

The proposed network architecture appears on Figure (1). From a qualitative point of view, the graphical component represents query, index terms, documents as nodes and the (in)dependence relations existing between nodes as arcs. Document and query nodes have binary domains. A document D_j is invoked or not, taking its values in the domain $\{d_j, \overline{d_j}\}$. The activation of a document node, i.e. $D_j = d_j$ (resp. $\overline{d_j}$) means that a document is relevant or not. A

query Q takes its values in the domain $\{q, \overline{q}\}$. As only the query instantiation is of interest, we consider $Q = q$ only, and denote it as Q. The domain of an index term node T_i, is $\{t_i, \overline{t_i}\}$. $(T_i = t_i)$ refers to the presence of a term in a query or a document and thus is *representative* of the document or a query to a certain degree. A *non representative* term, denoted by $\overline{t_i}$ is a term absent from (or not important in) the object.

Let $\mathcal{T}(D_j)$ (resp. $\mathcal{T}(Q)$) be the set of terms in-

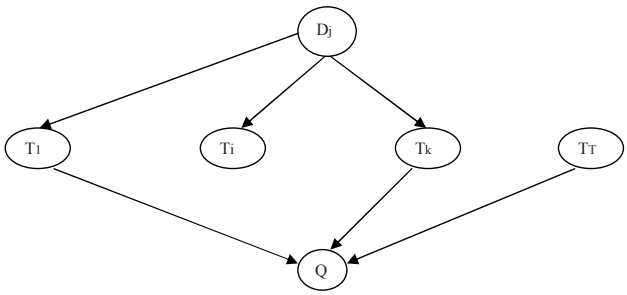

Figure 1: Model architecture

dexed in document D_j (resp. in the query). The query expresses the request for documents containing some terms but excluding other terms. Arcs are directed from document node to index term nodes defining dependence relations existing between index terms and documents. The architecture of the network depends on the terms appearing in the query and the document. The query instantiation only gives evidence to propagate through invoked terms thus, arcs are directed from term to query nodes. The terms appearing in the user query form the parent set of Q in the graph. There is an instantiation of the parent set $Par(Q)$ of the query Q that represents the query in its most demanding (conjunctive) form. Let θ^Q be such an instantiated vector. Any instance of the parent set of Q is denoted θ. We show, later in this section, how values are assigned to arcs given these domains.

2.2 Query propagation

One main original idea behind our possibilistic model concerns the relevance interpretation. Instead of using a unique relevance value of a document with respect to a query, we propose a possibilistic approach [5] to compute relevance. This model should be able to infer propositions like:

- It is plausible to a certain degree that the document is relevant for the user need.
- It is almost certain (in possibilistic sense) that the document is relevant to the query.

The first kind of proposition is meant to eliminate irrelevant documents (weak plausibility). The second answer focuses attention on what looks very relevant. Under a possibilistic approach, given the query, we are thus interested in retrieving necessarily or at least possibly relevant documents. The query evaluation consists in the propagation of new evidence through activated arcs to retrieve relevant documents. The propagation process evaluates the following quantities:

$$\Pi(d_j \mid Q) = \frac{\Pi(Q \wedge d_j)}{\Pi(Q)}, \quad \Pi(\overline{d_j} \mid Q) = \frac{\Pi(Q \wedge \overline{d_j})}{\Pi(Q)},$$

and $N(d_j \mid Q) = 1 - \Pi(\overline{d_j} \mid Q)$. The possibility of Q is $\Pi(Q) = max(\Pi(Q \wedge d_j), \Pi(Q \wedge \overline{d_j}))$. Given the model architecture, $\Pi(Q \wedge D_j)$ is of the form:

$$\max_{\theta}(\Pi(Q \mid \theta) \cdot \prod_{T_i \in \mathcal{T}(Q) \wedge \mathcal{T}(D_j)} \Pi(\theta_i \mid D_j) \cdot \Pi(D_j) \cdot \prod_{T_k \in \mathcal{T}(Q) \setminus \mathcal{T}(D_j)} \Pi(\theta_k))$$

for θ being the possible instances of the parent set of Q and for $D_j \in \{d_j, \overline{d_j}\}$.

3 Query representation

The possibility of the query given the index terms depend on query interpretation. Several interpretations exist, whereby query terms are aggregated by *conjunction*, *disjunction*... or, like in Bayesian probabilistic networks, by *sum* and *weighted sum* as proposed for example in the works of Turtle [14]. The basic idea is that for any instantiation θ, the conditional possibility $\Pi(Q \mid \theta)$ is specified by some aggregation function merging elementary possibilistic likelihood functions $\Pi(Q \mid \theta_i)$ where θ_i is the instance of T_i in θ. Each $\Pi(Q \mid \theta_i)$ is the weight of instance θ_i in view of its conformity with the instantiation of T_i in the query (in θ^Q). We do not consider relations that may exist between terms. Indeed, it would be difficult (space and time consuming) to store all possible query

term configurations or to compute them when the query is submitted to the system. A reasonable organization is to let each query term holds separately the weights associated to the query. When the user does not give any information on the aggregation operators to be used, the only available evidence one can use is the importance of each query term in the collection. This evidence is available for single terms that have to be combined.

3.1 Boolean aggregations

For a Boolean *AND* query, the evaluation process searches documents containing all query terms. Then, $\Pi(Q \mid \theta_i) = 1$ if $\theta_i = \theta_i^Q$, and 0 otherwise. The possibility of the query Q given an instance θ of all its parents, is given by $\Pi(Q \mid \theta)$ where $\Pi(Q \mid \theta) = 1$ if $\forall T_i \in Par(Q)\ \theta_i = \theta_i^Q$ means that the term T_i in θ is instantiated as in the query and 0 otherwise. Generally this interpretation of the query is too demanding.
For a Boolean *OR* query, the document is already somewhat relevant if there exists a query term in it. The pure disjunctive query is handled by changing $\forall$ into $\exists$ in the conjunctive query. But this interpretation is too weak to discriminate among documents.

3.2 Quantified aggregation

The final document relevance increases with the number of present query terms. Assume a query is considered satisfied by a document if they have at least K common terms. Consider an increasing function, $f(\frac{K(\theta)}{n})$, where $K(\theta)$ is the number of terms in the query instantiated like in a given configuration θ of $Par(Q)$, given that the query contains n terms. It is supposed that $f(0) = 0$ and $f(1) = 1$. For instance, $f(i/n) = 1$ if $i \geq K$, and 0 otherwise requires that at least K terms in the query are in conformity with θ. But more generally f can be a non-Boolean function (a fuzzy quantifier [15]). The possibility of the query Q given an instance θ of all its parents, is given by:

$$\Pi(Q \mid \theta) = f(\frac{K(\theta)}{n}) \tag{1}$$

3.3 Noisy OR

In general, we may assume that the conditional possibilities $\Pi(Q \mid \theta_i)$ are not Boolean-valued, but depend on suitable evaluations of terms t_i. A possible query term combinations can be "noisy-OR" [6] based. It means that $\Pi(Q \mid \theta)$ is evaluated in terms of conditional possibilities of the form $\Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k})$ using a probabilistic sum. In this case, under the Closed World Assumption (CWA), $\Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k})$ is considered to be the same as $\Pi(Q \mid t_i)$ and θ^Q contains only positive $t_i's$. In the following, this is what is assumed, i.e. $\theta_i^Q = t_i \forall T_i \in Par(Q)$. However, the possibility network approach could account for the presence of negative terms in the query.

4 Weighting schemes

For the evaluation process, we have to find an effective weighting scheme for assessing the possibility degrees of arcs existing between pair of term-query and document-term nodes. In the first part of this section, we define weights assigned to document-term arcs then to term-query arcs (root terms). We at least define weight to aggregate query and to prior possibility of documents.

4.1 Index document terms weighting

Our approach tries to distinguish between terms which are possibly representative of documents (whose absence rules out a document) and those which are necessarily representative of documents, i.e. terms which suffice to characterize documents. One possible convention is :
Postulate 1: A term appearing in a document is more or less certainly representative of that document;
Postulate 2: A document is all the more necessarily suggested by a term as the latter appears with higher frequency in that document and with lower number of apparitions in the rest of the whole collection.
Arcs from documents to terms are evaluated on the basis of probability masses proposed by Dempster Shafer theory [10]. Masses are assigned to elementary propositions and to disjunctions thereof. The frame of discernment of a term is

$\Theta_i = \{t_i, \overline{t_i}\}$. Probability masses are assigned to subsets $\{t_i\}, \{\overline{t_i}\}, \{t_i, \overline{t_i}\}, \emptyset$ meaning that the term is, respectively, surely representative, surely not representative, or its representativeness is unknown or yet conflicting. The basic probability assignment (bpa), denoted by m, is a function such as:

$$m(\emptyset) = 0, \quad m(t_i) + m(\overline{t_i}) + m(\Theta_i) = 1 \qquad (2)$$

The probability masses, in the context of $D_j = d_j$, are given by:

$$m(\{t_i\} \mid d_j) = ntf_{ij},\ m(\{t, \overline{t}\} \mid d_j) = 1 - ntf_{ij},$$

where ntf_{ij} is normalized term frequency, $ntf_{ij} = \frac{tf_{ij}}{max_{t_k \in d_j}(tf_{kj})}$.

It means that a term present in the document, is certainly representative of that document at least to degree ntf_{ij}. The second mass $1 - ntf_{ij}$ can freely move to any element of Θ_i. In our context, we are interested in defining conditional possibility degrees of representativeness which are obtained respectively from the mass functions. We define the conditional possibility $\Pi(T_i \mid d_j)$ as a Shafer plausibility function due to consonance of the mass function[1]. In the context $D_j = d_j$, we obtain:

$$\Pi(t_i \mid d_j) = 1;\ \Pi(\overline{t_i} \mid d_j) = 1 - ntf_{ij}$$

Then $1 - ntf_{ij}$ represents how sure an unfrequent term t_i in d_j is irrelevant to d_j.
An important term in a collection is a term which appears with high frequency in few documents of the collection, but not too rare as well explained in [13]. We assume that the certainty of retrieving a relevant document by means of a term is related to the importance of this term in the whole collection. The importance of a term t_i for retrieving a document j is usually taken as $\phi_{ij} = \frac{log\frac{N}{n_i}}{log(N)} \cdot ntf_{ij}$ where N is the number of documents in the collection and n_i is the number of documents containing term t_i.
When we are in the context of $D_j = \overline{d_j}$, we define a bpa, by:

$$m(\{\overline{t_i}\} \mid \overline{d_j}) = \phi_{ij},\ m(\{t_i, \overline{t_i}\} \mid \overline{d_j}) = 1 - \phi_{ij}$$

[1]Another assumption can be that ntf_{ij} represents how possibly relevant is a term for a document, as done in [3]

ϕ_{ij} is interpreted as the extent to which, if d_j is not a good document, we should not use term t_i. If t_i is not relevant to d_j then the fact that d_j is not a good document leaves us free to use t_i or not. As previously, we get: $\Pi(\overline{t_i} \mid \overline{d_j}) = 1$
$\Pi(t_i \mid \overline{d_j}) = 1 - \phi_{ij}$. Table 1 gives the conditional possibilities.

Table 1: Conditional possibility $\Pi(T_i \mid D_j)$

	d_j	$\overline{d_j}$
t_i	1	$1 - \phi ij$
$\overline{t_i}$	$1 - ntf_{ij}$	1

4.2 Root terms weighting

Root term nodes are terms present in the query and absent from the document. Weights are assigned to them are useful to decrease the final document relevance. Those weights may either set to a constant α or to maximum weights assigned in the context of the treated document. The more important a query term, t_i, absent from a document is, the smaller α_i is, and conversely. The importance of a term is reflected by its density distribution over the collection. Instead of the *nidf* ($nidfi = log(\frac{N}{n_i})$) coefficient, a more refined discrimination factor is $df_{t_i} = -\sum_j p_{ij}\, log\, p_{ij}$, where $p_{ij} = \frac{\frac{tf_{ij}}{l_j}}{sum_{k=1,..,N}\frac{tf_{ik}}{l_k}}$ is used. Thus, $\Pi(\theta_k) = 1$ if $\theta_k^Q = \overline{t_k}$ and $\Pi(\theta_k) = df(t_k)$ if $\theta_k^Q = t_k$.

4.3 Prior possibility of documents

In absence of information, the *a priori* possibility of a document node is uniform (= 1). Actually, we can obtain information on a document given the importance of its terms, its length etc. This knowledge can be given for instance, by the user, the user profile defined on the collection etc. Hence, for example, if we are interested in retrieving long documents, we define the prior possibility of document $D_j = d_j$, $\Pi(d_j) = \frac{l_j}{max_{k=1,..,n} l_k}$ where l_j is the length in frequency of document d_j; $l_j = \sum_i tf_{ij}$. The shorter the document, the less possibly relevant it is.

4.4 Conditional possibility for query node

The primitive terms in a noisy OR are $\Pi(Q \mid t_i \wedge_{k \neq i} \overline{t_k}) = nidf_i$, denoted $1 - q_i$ for simplicity. Then

$$\Pi(Q \mid \theta) = 0 \text{ if } \nexists \text{ i s.t. } \theta_i = \theta_i^Q$$
$$= \frac{1 - \prod_{i:t_i=\theta_i=\theta_i^Q} q_i}{1 - \prod_{T_k inPar(Q)} q_k} \text{ otherwise}$$

It should be remembered that Q contains only positive terms, i.e. $\theta_i = t_i$.

5 Example

Assume a collection composed of 5 documents:

$D_1 = \{4t_9, 6t_4\}, D_2 = \{3t_2, 10t_3, 15t_5, 6t_6, 10t_7, 12t_8\},$
$D_3 = \{t_1, t_2, t_5\}, D_4 = \{t_1, 15t_3, t_4\},$
$D_5 = \{15t_1, 15t_2, 15t_3\}$
$Q = \{t_1, t_3, t_6\}$

Taking document D_1 as illustration, the terms t_9 and t_4 appears respectively 4 and 6 times in document D_1. As defined in the previous section, the different weights for query terms, in the context of their parents, are given in the table below. *A priori* possibility, for instantiated document $D_j = d_j$, computed as defined in section 4.3, is given in the table 3. Weights assigned to root terms depending on their density as defined in section 4.2 are given in table 4. A constant value (equals to 0.1) is added to df_i to avoid 0 value[2].

Table 2: Weights assigned to query terms t_1, t_3, t_6

	D_3	D_4	D_5
$1 - \phi_{t_{1j}}$	0.682	0.978	0.682
	D_2	D_4	D_5
$1 - \phi_{t_{3j}}$	0.788	0.682	0.682
	D_2		
$1 - \phi_{t_{6j}}$	0.6		

Table 3: A priori possibility

	β_{d_j}
d_1	0.18
d_2	1
d_3	0.05
d_4	0.30
d_5	0.80

Table 4: Root term weights

df_{t_1}	1
df_{t_3}	0.98
dft_6	0.2

Assuming the query aggregated by a noisy OR, we are interested in documents having at least one common term with the query. Thus :

$\Pi(Q \wedge d_2) = 0,828; \Pi(Q \wedge \overline{d_2}) = 0.39$
$\Pi(d_2 \mid Q) = 1; \Pi(\overline{d_2} \mid Q) = 0,47; N(d_2 \mid Q) = 0,53$
$\Pi(d_3 \mid Q) = 0.287; \Pi(\overline{d_3} \mid Q) = 1; N(d_3 \mid Q) = 0$
$\Pi(d_4 \mid Q) = 0.3; \Pi(\overline{d_4} \mid Q) = 1; N(d_4 \mid Q) = 0$
$\Pi(d_5 \mid Q) = 0.430; \Pi(\overline{d_5} \mid Q) = 1; N(d_5 \mid Q) = 0$

Documents are retrieved given their necessity measures and if 0 given their possibility measures. If the possibility degree of a document given the query equals 0, the document is discarded. Thus, for the example above, documents are retrieved in order D_2, D_5, D_4, D_3. This ranking is not surprising given that the top ranked document, D_2, contains the term having the most discriminative power. Besides this document is the longest. Documents D_5, D_4 contain the same terms but document D_5 is longer than document D_4. Although t_3 has more power to discriminate between documents and it is more dense in document D_4 than in D_5, the preference on document D_5 (ranked better than D_4) can be explained by the length impact on ranking. The last ranked document is D_3 and this is due to the fact that it is the shortest document and that there is no preference on representativeness of terms (t_1 and t_3 have same frequency inside it).
A comparison with vector space, probabilistic and inference networks models shows that we retrieve documents in the same ranking as the inference network model. The used weighting scheme on term documents for inference networks

[2]The more df_i of a term t_i is high in table 4, the more t_i absence does not affect the document relevance.

is $0.5+0.5*ntf*nidf$ [14]. The probabilistic model retrieves only the document D_2. Used weights for this comparison are those of $BM-25$ [7][8]. The vector space model using pivot length normalization retrieves document D_3 before document D_4, which is the main difference with inference network or possibilistic models. For comparison with the vector space model we use weights and evaluation scheme of pivot normalization length [11].

6 Conclusion

This paper presents a new IR approach based on possibility theory. In a general way, the possibility measure is convenient to filter out documents (or index terms from the set of representative terms of documents) whereas necessity reinforces document relevance (or index terms representativeness). First experiments on real collections indicate the proposed approach is very promising [4].

References

[1] S. Benferhat, D. Dubois, L.Garcia, H. Prade: Possibilistic logic bases and possibilistic graphs. In Proc. of the 15th Conference on Uncertainty in Artificial Intelligence, (1999) 57-64.

[2] C. Borgelt, J. Gebhardt and R. Kruse: Possibilistic graphical models. Computational Intelligence in Data Mining, CISM Courses and Lectures 408, Springer, Wien, (2000) 51-68.

[3] A.H. Brini, M. Boughanem and D. Dubois (2004). "Towards a Possibilistic Approach for Information Retrieval". Proc. EUROFUSE Data and Knowledge Engineering, Warsaw, pp 92-102. (2004).

[4] A.H. Brini, M. Boughanem and D. Dubois (2005). "A Model for Information Retrieval based on Possibilistic Networks". Submitted to SPIRE 2005.

[5] D. Dubois H. Prade Possibility Theory, Plenum, 1988.

[6] Judea Pearl (1988), Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann San Mateo, Ca., 1988.

[7] S.E. Robertson & S. Walker. "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval". In Proc. of the 17th Ann. Int. ACM SIGIR Conf. on Research and Dev. in Information Retrieval, Springer-Verlag. 1994, pp 232-241.

[8] S.E Robertson, S. Walker, S. Jones, M.M Hancock-Beaulieu & M. Gatford. "Okapi at TREC-3". In Proc. of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-225, 1995, pp 109-126.

[9] G. Salton. "The Smart retrieval system-experiments". In Automatic Document Processing, Prentice Hall Inc. 1971.

[10] Shafer, G., (1976). A mathematical theory of evidence, Princeton Univ. Press.

[11] A. Singhal, C. Buckley, M. Mitra. "Pivoted document length normalization". In Proc. of the ACM-SIGIR Conf. on Research and Development in Information Retrieval. 1996, pp 21-29.

[12] K. Sparck Jones, S. Walker and S.E. Robertson. " A probabilistic model of information retrieval: development and comparative experiments". Parts 1 & 2". IPM, 36(6), 2000, 779-808 and 809-840.

[13] K. Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". Journal of Doc, 28, 1972, 111-21.

[14] H.R. Turtle and W.B. Croft (1990). "Inference networks for document retrieval". In Proc. 13th Int.Conf. on Research and Development in Information Retrieval, 1990, pp 1-24.

[15] R. R. Yager and H. Legind Larsen. "Retrieving information by fuzzification of queries". Int. Jour. of Intelligent Inf. Systems, 1993, Vol. 2(4).