International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 8 (2016) pp. 226-234
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Fuzzy Networks based Information Retrieval Model

Kamel Garrouch[1], Mohamed Nazih Omri[2]

[1] Faculty of Science Monastir, University,
Avenue de l'environnement, Monastir 5019 , Tunisias
Kamelg_2001@yahoo.fr

[2] Faculty of Science, Monastir, University,
Avenue de l'environnement,, Monastir 5019 , Tunisias
MohamedNazih.Omri@fsm.rnu.tn

***Abstract*: We describe an Information Retrieval Model based on fuzzy Networks that incorporates dependence relationships between indexing terms. We details the design and implementation of a new Information Retrieval Model based on Fuzzy Network. From this Network, most relevant term to term dependence relationships are extracted using within document terms dependency analyses. The criteria used to select these dependence relationships are the strength of dependency of each pair of terms within each document and the strength of dependency of each pair of terms in the entire document collection. The relevance of a document to a query is interpreted by two degrees: the necessity and the possibility. The necessity degree evaluates the extent to which a document is relevant to a query, whereas the possibility degree evaluates the reasons of eliminating irrelevant documents. These two measures are also used for quantifying terms-terms links and terms-documents links. Experiments carried out on three standard document collections show the effectiveness of the model.**

Keywords: Information retrieval, Possibility theory, Possibilistic networks, Term dependency

I. Introduction

The field of Information Retrieval (IR) has been defined by Salton [26] as the subject concerned with the representation, storage, organization, and accessing of information items. The main objective of an Information Retrieval System (IRS) is to select, among a large collection of documents, those that are relevant to a user's query. Given a document collection, the first step of an information retrieval process is to create a representation for each document in a suitable form to be managed by a computer. This task is called indexing which is generally done off line. The result of this task is a set of terms extracted from each document that should appropriately express its content. Because these terms are not equally important, they can be weighted in a second step in order to highlight their importance in the documents to which they belong. A weighted indexed document could be $D_j = \{(t_{l\,j}, w_{lj}), ..., (t_{k\,j}, w_{kj})\}$), where each w_{ij} is the weight associated to the corresponding term t_i. Usually, we use the weight known as tf*idf weight. In this case, the weight value associated to a given term t_i in a given document d_j is computed by multiplying the frequency of the term in this document (tf_{ij}) by its inverse document frequency (idf_i) in the document collection (i.e. $w_{ij} = tf_{ij}$ x idf_i) .

In order to produce a representation for the user's query, the latter is also indexed in a third step. This task is done on line. Once the query representation is produced, it can be matched to documents representations, in a forth step, in order to retrieve documents relevant to the query. The matching task is based on functions implemented by the corresponding information retrieval model. Actually, a similarity score between the user query and each candidate document is computed using a scoring mechanism. The result of this stage is a ranking of documents sorted by their proximity to the query.

Information retrieval systems are based on different theoretical models, which determine how indexing and matching tasks are conducted. The most prevalent models are Boolean, Vector Space, Probabilistic, and Language Modeling. Existing Information Retrieval Models (IRM) can be classified into two families. The first is based on the assumption that terms indexing one document are statistically independent [3], [7], [8], [18], [21], [22], [27]. This assumption is made because of the great expense that is expected to be incurred if term dependences are used [17], [25]. Thus, it makes the retrieval easier to be implemented. The second considers that the previous assumption is obviously wrong. The main idea here is that terms dependency is an indispensable consequence of language use, since words are not actually independent [20]. Therefore, independence assumption may cause a loss of information corresponding to the term dependence relationships, which may lower the performance of information retrieval systems. Consequently the use of term dependencies can improve the performance of IRS [28].

Most of models breaking the independence assumption are probabilistic. Some of them make an explicit graphical representation of term dependencies using a Bayesian network [10], [11], [12], [14], [16], [17], [19], [28]. Others integrate term dependencies in their matching mechanism between documents and query representations [2], [5].

In this paper, we are interested in information retrieval models that break the independence assumption and make an explicit representation of terms dependence relationships. Several information retrieval models of this kind have been proposed. However, to extract dependent pairs of terms, these models generally use a formula that analyzes terms co-occurrence between each pair of terms in the whole documents collection. This method has three problems. The first is that it doesn't consider the strength of terms dependencies inside each document. The second is that it ignores vagueness and fuzziness which is inherent to natural language. The third is that the used formula leads to a great number of linked terms and to weak values of dependencies [20].

To overcome these problems, this paper introduces a new possibilistic network based information retrieval model. The aims of this model are: 1) to extract most relevant term to term dependency relationships, by the means of within document terms dependency analyses. 2) to use possibilistic measures for both the quantification of the relevance of a document to a user query and the quantification of the strength of dependency relationships between pairs of terms.

The paper is organized as follows. In section 2, we briefly present different approaches to IR using Bayesian networks. In section 3, we briefly introduce the Possibilistic network background needed to understand the rest of the paper. In section 4, we describe in detail the proposed model. Section 5 shows experiments carried out on three document collections. The final section presents conclusions and future lines of research.

II. Related Works

Information Retrieval models that integrate terms dependencies are facing two main problems. The first is how to obtain terms dependency relationships efficiently, and the second is how to use them to retrieve documents, given a user query [28]. Most of the models that make an explicit representation of indexing term dependencies are based on Bayesian networks. In order to solve the efficiency problem, caused by the great expense incurred if higher order dependencies are used in estimating probabilities, these models are generally based on two main simplifying restrictions:

1. Fixed dependence relationships: the structure of the model, encoding the dependence relationships between variables, is set a priori, without considering any potential knowledge that might be mined from the collection.
2. Simplified estimation of conditional probabilities distribution: in order to avoid the large space needed to store all the probabilities relevant to the process, these models make use of canonical models to do this task.

Based on these simplifications, several models have been proposed. The main differences between them are the number of subnetworks composing the Bayesian network, the process used to make the orientation of arcs and the modelling of the (in)dependence relation between term nodes. We are briefly going to review some works which are based on this kind of model.

De Campos et al [10] proposed a Bayesian Network Retrieval Model (BNRM) composed of two different subnetworks: the term subnetwork and the document subnetwork. The former’s nodes represent indexing terms. Nodes links are used to depict dependence relationships between indexing terms. The latter’s nodes represent the set of documents. The relationships between a document and its indexing terms are presented by the links between the two subnetworks. In this model there is no node for the user’s query. In fact, query terms are considered as evidence that should be introduced into the system. To reduce the computation cost, this model uses canonical models instead of learning algorithms to estimate the conditional probability distribution of nodes. Canonical models are also used to do inference process. In a similar approach, Dongyu et al [12] proposed a model having almost the same as the BNRM’s one. However, the set of arcs are not oriented in the term subnetwork. This constitutes the main difference between the two models. A third model, proposed by de Campos et al [11], uses two term-layers to encode term relationships. It is based on the use of a term clustering technique to extract the strongest relationships among terms. Therefore, the complete Bayesian network contains three simple layers: two term layers and a document layer. A fourth model with two terms, layers was proposed by Xu et al [28]. Here, the term relationships are mined by using word similarity extracted from a thesaurus.

The above models have two advantages: 1) they incorporate terms dependence relationships, 2) they reduce computation cost by using some simplifications, such as the fixed structures and the canonical models. However, for these models, term dependencies extraction procedure is not really based on a within document analyses. In fact, they use a formula that analyses terms co-occurrence between each pair of terms in the whole documents collection in order to quantify the degree to which two terms are considered as dependant without taking into consideration the strength of their dependence relationship within each document. This may leads both to a great number of linked terms and to weak values of dependencies [10].

Possibilistic Network based Information Retrieval Models (PNIRM) are not as numerous as their Bayesian network counterpart. Actually, the first important one was proposed by Boughanem et al [3]. In this model, the relevance assessment of a document to user a query is based on two possibilistic measures: possibility and necessity. The necessity degree evaluates the extent to which a given document is relevant to a query, whereas the possibility degree evaluates the reasons of eliminating irrelevant documents. The advantage of this model is that it deals with the concept of relevance under a possibilistic framework. However, it is based on the independence assumption between indexing terms. Based on this possibilistic interpretation of the relevance, Garrouch et al [14] proposed to combines the advantages of Bayesian networks based information retrieval models with the possibilistic network models described above. This proposition was putted in practice in [16] where a PNIRM integrating terms dependencies was developed. Although this model was the first of its kind, it showed a weak retrieval performance.

III. Possibility theory and Possibilistic Networks

Introduced by Zadeh [29] and developed by Dubois and Prade [13], possibility theory constitutes a powerful and simple alternative to probability theory in particular for dealing with some types of uncertainty [15].

A. Possibility theory

The basic concept in the possibility theory is the notion of possibility distribution denoted by π which is a mapping from the universe of all possible states of the world Ω (universe of discourse) to the unit interval [0, 1]. π(w) evaluates the plausibility that w is the actual value of some variable to which π is attached. π(w) = 0 means that w is impossible. π(w) =1 means that w is completely possible (unsurprising).
Two dual measures are used in possibility theory: the possibility measure Π(A) and the necessity measure N(A). The possibility of an event A, noted Π(A) describes the most normal situation in which A is true. It is defined by

$$\Pi(A) = \max_{w_i \in A}(\pi(w_i)) \quad (1)$$

The necessity of an event A reflects the most normal situation in which A is false. It is defined by:

$$N(A) = \min_{w \notin A}(1-\pi(w_i)) = 1-\Pi(\bar{A}) \quad (2)$$

The distance between N(A) and Π(A) evaluates the level of ignorance on A [3].

A second concept in the possibility theory is the notion of possibilistic conditioning. It consists of updating the current beliefs encoded by the possibility distribution π by the arrival of a new sure piece of information ∅⊆Ω. In possibility theory there are two possible definitions of conditioning; one is based on the product and the other on the minimum. In this work, given a normalized possibility distribution π, we focus only on the product-based conditioning defined as follows:

$$\pi(w_i|\varphi) = \begin{cases} \dfrac{\pi(w_i)}{\Pi(\varphi)} if \;\; w_i \in \varphi \\ 0 \quad otherwise \end{cases} \quad (3)$$

B. Product based possibilistic networks

A product-based possibilistic network over a set of variables V={A$_1$, A$_2$,…, A$_N$} is a directed possibilistic graph where conditionals are defined using product-based conditioning. It is characterized by a qualitative component and a quantitative component. The first one is a directed acyclic graph which encodes independence relation sets. The second component quantifies the strength of distinct links of the graph and consists of a set of conditional possibility tables of each node in the context of its parents. These possibility distributions should respect the following normalization rules:

For each variable V

- If V is a root node and dom(V) is the domain of V, the prior possibility of V have to satisfy:
$$\max_{v \in dom(v)} \Pi(v) = 1$$
- If V is not a root node, the conditional distribution of V in the context of its parents should satisfy:

$$\max_{v \in dom(v)} \pi(v|Par_v) = 1, \quad Par_v \in dom(Par_v)$$

where *Par$_v$* is a configuration of parent variables of V and dom *(Par$_v$)* is the Cartesian product of domains of parents of the variable V.
The possibility distribution of product-based possibilistic networks, π$_p$, obtained by the chain rule is:

$$\pi(V_1,\ldots,V_n) = \prod_{i=1} \pi(V_i \,|Par_{v_i}) \quad (4)$$

IV. Proposed model

In this work, a new Possibilistic Network based Information Retrieval Model (PNRIM) is presented. It differs from existing proposals on three main points:
1) The quantification of the strength of dependency relationships between terms is based on two possibilistic measures (possibility and necessity). The possibility of dependence of a pair of terms, denoted by ($\Pi_{dep}(t_i,t_j)$), is meant to eliminate irrelevant dependencies. Actually, if $\Pi_{dep}(t_i,t_j) = 0$, it is certain that the two terms are not dependent. However $\Pi_{dep}(t_i,t_j) =1$ does not imply that the pair of terms are dependent, only that nothing prevents them from being dependent. The necessity of dependence of a pair of terms denoted by ($N_{dep}(t_i,t_j)$), focuses attention on relevant dependencies. Since $N_{dep}(t_i,t_j) >0 \Rightarrow \Pi_{dep}(t_i,t_j) = 1$, only possibly dependent pairs of terms can be considered as necessarily dependent (to a certain degree). Under a possibilistic approach, given a set of document, we are interested in extracting necessarily dependent pairs of terms; or at least possibly dependent ones.
2) These possibilistic measures are also used for the relevance quantification of a document to user query. The possibility of relevance of a document to a user query denoted by ($\Pi(d_j|Q)$), is meant to eliminate irrelevant documents. If $\Pi(d_j|Q) = 0$, it is certain that the document d$_j$ is not relevant to the query Q. However, $\Pi(d_j|Q)=1$ does not imply that the document is relevant, only that nothing prevents the document from being relevant. The necessity of relevance of a document to a user query denoted by ($N(d_j|Q)$), focuses attention on relevant documents. Since $N(d_j|Q)>0 \Rightarrow \Pi(d_j|Q)=1$, only possibly relevant documents can be considered as necessarily relevant (to a certain degree). Thus, the proposed model will be used to retrieve necessarily relevant documents or at least possibly relevant ones.
3) The approach proposed for the extraction of the set of dependent pairs of terms, from a given document collection is based on two criteria. The first one is the strength of

dependency of each pair of term within each document. The second criterion is the strength of dependency of each pair of terms in the entire document collection. The hypothesis used here is : the greater is the number of documents where a pair of terms co-occurs, the greater is the belief about their dependency.

A. *The structure of the model*

The structure of the proposed model is composed of two layers: a term layer and a document layer (Figure.1). The first one contains the set of indexing terms T = {T$_i$, i = 1… M}, M being the number of terms used to index the document collection. The domain of an index term node T$_i$ is $\{t_i, \bar{t}_i\}$.

T$_i$ = t$_i$ refers to the fact that the term is selected to represent a document. A non representative term, denoted by $\bar{t}_i$ is a term absent from (or not important in) the object. A link between two term nodes means that they are dependent.

The topology for representing term to term relationships that supports the model is a polytree. It is adopted because it represents a good alternative for managing domains with a large number of variables, such as information retrieval applications, where we have to deal with thousands of terms or concepts, and each one represents a variable [9].

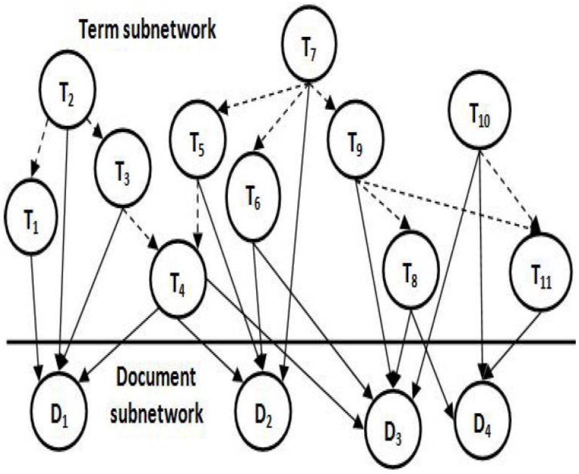

Figure 1. Structure of PNIRM

The second layer is the document layer. It contains the set of documents nodes D = {D$_j$, j=1... N}, with N being the number of documents in the collection. The domain of a document node is $\{d_j, \bar{d}_j\}$. D = d$_j$ means that the document is relevant to a given query. $D_j = \bar{d}_j$ means that the document is not relevant for a given query (i.e. it does not satisfy the user's information need). Arcs are directed from terms nodes to their corresponding documents nodes. There are no links joining the document nodes between them.

B. *Model building*

The general principle used by structure learning algorithms to build the structure of the polytree based network from empirical data is composed of two tasks: the first one is to construct an undirected graph whose edges connect every pair of variables which are not independent. The second task is to give a direction to each of these edges [9], [24]. Based on this principle, the process used in our model to build the structure of the network consists of four steps. The first step is to discover the set of most relevant term dependencies from the set of documents. The second step is to build the tree skeleton using a greedy algorithm in order to obtain a Maximum Weight Spanning Tree (MWST). The aim of these two steps is to construct the undirected graph. The third step deals with the orientation of the edges in the tree to make up a polytree. The last step is to join each indexing term to its corresponding document.

1) *Construction of the list of dependencies*

We propose to keep in our model only the most relevant terms dependence relationships. To reach this goal, we propose a process that: 1) investigates the collection of documents one by one, 2) creates a list of most relevant terms dependence relationships from each document and 3) merges the obtained lists together, in order to have a final list of term dependencies. Two complementary measures are used to quantify the dependence relationships within one document: the possibility of dependence and the necessity of dependence. The first measure describes to which extent two terms t$_i$ and t$_j$ are possibly dependent within one document. We assume that two terms are possibly dependent if their within document frequency meet or exceed a specified minimum value (β). To assess the possibility of dependence between two terms t$_i$ and t$_j$ within one document d$_k$, we propose to use the following formula:

$$\Pi_{dep_{dk}}(t_i, t_j) = \begin{cases} 1 & if\ \ tf_{ijk} > \beta \\ 0 & otherwise \end{cases} \qquad (5)$$

The aim of the use of this formula is to focus on the list of possibly dependent pairs of terms. Thus, all pairs of terms having a possibility value of dependence bellow β are excluded. At present, there is no well-motivated basis for selecting values of β that can be expected to yield good retrieval results for a particular document collection. Thus, optimal values of this threshold must be determined empirically for each collection. A large number of experiments testing several values for the threshold have been conducted. Details about these experiments are specified in section 5.

The necessity of dependence describes the certainty degree of terms dependencies. We assume that two terms t$_i$ and t$_j$ are necessarily dependent to a certain degree within one document d$_k$ if they are at least possibly dependent. To assess the necessity degree of dependence of two terms t$_i$ and t$_j$ within one document d$_k$, we propose to use the following formula:

$$N_{dep_{dk}}(t_i, t_j) = \frac{tf_{ijk}}{\max\limits_{d_l \in D}(tf_{ijk})} \qquad (6)$$

With tf$_{ijk}$ is the measure of the co-occurrence frequency of the terms t$_i$ and t$_j$ in the document d$_k$ and $\max\limits_{d_l \in D}(tf_{ijk})$ is the maximum value of co-occurrence of the terms t$_i$ and t$_j$ in the document collection. It is used as a normalisation factor.

The choice of this formula is based on the assumption that terms that occur frequently together in a document are generally about the same subject. Thus, term co-occurrence data obtained from the analysis of each document can be used to identify some of the semantic relationships that exist between terms [4]. In other words, the formula quantifying the strength of the dependency of a pair of terms has to be based on the information about their co-occurrence frequency.

The formulas 5 and 6 allow us to extract, from each document, a list of dependent pairs of terms. Some of them exist only in one document, while others exist in several documents. Consequently, we have to transform the set of lists obtained from each document to a single one. Our idea is to strengthen the values of dependency of pairs of words that occur in many documents. Thus, the final values of dependency of a given pair of terms will be obtained by adding up the set of dependency values obtained from each document, while taking into account the number of documents in which they co-occur. Therefore, to quantify the strength of the dependence relationship between two terms in the whole document collection, we propose to use two possibilistic measures: the first is the possibility of dependence, used in order to eliminate pairs of terms that are not possibly dependent in the document collection. It is described as follow:

$$\Pi_{dep_{coll}}(t_i,t_j) = \begin{cases} 1\, if\ \exists\, d_k \in coll \setminus N_{dep}(t_i,t_j) > 0 \\ 0\quad otherwise \end{cases} \tag{7}$$

The second measure is the necessity of dependence. It describes to which extent two terms t_i and t_j are necessarily dependent in the document collection. Its value is obtained gradually from N documents. The proposed formula is as follows:

$$N_{dep_{coll}}(t_i,t_j) = \frac{n_{ij}}{N} \times \frac{\sum_{i=1}^{n_{ij}} N_{dep_{dk}}(t_i,t_j)}{N} \tag{8}$$

Here,

$N_{dep_{dk}}(t_i,t_j)$ is the necessity dependence measure of a pair of terms t_i and t_j in a document d_k,

N is the number of document in the collection,

n_{ij} is the number of documents containing both t_i and t_j and k have to verify $1 \leq k \leq N$.

2) Construction of the tree skeleton

The objective of this step is to build the tree skeleton of the model which is a Maximum Weight Spanning Tree (MWST), i.e. a tree that maximizes the sum of its links weights. The idea here is to preserve the edges having the strongest dependency relations, with the restriction that the resultant structure must be singly connected [9]. To reach this objective, we first assume that the computed necessity values are link weights in the graph, and then we apply a greedy algorithm to get our MWST.

3) Orientation of the edges in the tree

Once the skeleton is built, the last part of the structure building process deals with the tree's orientation getting a polytree as a result. The aim here is to assign directions to each triplet's edges in the network. Since there are three possible types of orientation of triplets allowed in a polytree, the task is to choose the appropriate orientation for each triplet T_i—T_k—T_j.

Type 1: $T_i \rightarrow T_k \rightarrow T_j$,

Type 2: $T_i \leftarrow T_k \rightarrow T_j$,

Type 3: $T_i \rightarrow T_k \leftarrow T_j$

In a head to head pattern (Type 3), the instantiation of the node T_k should normally increase the degree of dependency between the variables T_i and T_j, whereas in a non-head to head pattern (Type 1 and Type 2), the instantiation of the node T_k should produce the opposite effect. Based on this idea, de Compos [9] proposes to compare the degree of dependency between T_i and T_j after the instantiation of T_k, $Dep(T_i,T_j|T_k)$ with the degree of dependency between T_i and T_j before the instantiation of T_k, $Dep(T_i,T_j|\varnothing)$ and direct the edges toward T_k, if the former is greater than the latter.

Logically, in the case of information retrieval, to have a head pattern, the three nodes, T_i, T_k, and T_j have to co-occur at least in one document; otherwise T_k can't have any influence on the dependency between T_i and T_j. In fact, if the link T_i—T_k is obtained by analyzing a document D_l and the link T_k—T_j is obtained by analyzing another different document D_m, then T_i and T_j are independent and can't be conditionally dependent to T_k. Thus, the triplet T_i—T_k—T_j can't be considered as head to head connection. In contrast, if the two links come from the same document then there is more chance that T_i and T_j are conditionally dependent to T_k.

Based on these facts, we propose to orient a triplet T_i—T_k—T_j as a head to head pattern only if the following two conditions are satisfied:

1. The three nodes T_i, T_k and T_j should co-occur in at least one document, and
2. $Dep(T_i,T_j|T_k) > Dep(T_i,T_j|\varnothing)$.

The first condition allows the reduction of the cost of computation time and storage needed for the orientation process. Indeed, given a set of triplets, only those satisfying the condition are concerned by the identification of the set of head to head patterns. Thus, the remaining triplets are discarded.

The dependency measure used in this case is based on the possibilistic mutual information [6] given by the following equations:

$$Dep(t_i,t_j|\varphi) = -\sum_{i,j} \pi(t_i,t_j)\log\frac{\pi(t_i,t_j)}{\min(\pi(t_i),\pi(t_j))} \tag{9}$$

$$Dep(t_i,t_j|t_k) = -\sum_k\sum_i\sum_j \pi(t_i,t_j,t_k)\log\frac{\pi(t_i,t_j,t_k)}{\min(\pi(t_i|t_k),\pi(t_j|t_k))} \tag{10}$$

After obtaining the set of head to head connections, we finish the orientation task by directing the remaining undirected edges without adding new head to head connections.

Once the polytree is learned, the last step to finish the structure construction is to join each term node with its corresponding document node. A link between a document node and a term node means that the term is chosen to represent the document. A missing link between them means that either the term does not exist in document or it can not represent the document.

C. Parameters estimation

Once the structure is built, the next step is to estimate the set of possibility distributions. In our model, we have three kinds of nodes: root term nodes, non root term nodes and leaf (document) nodes.

1) Root term nodes

For a root term node, we have to store the marginal possibility of relevance $\Pi(t_i)$ and the marginal possibility of being non-relevant $\Pi(\bar{t_i})$. Given that a weight w_{ij} is used in information retrieval field to assess the relevance of a given term t_i in a given document D_j, we think that it can also be used to measure the relevance of the term in the document collection. Thus the marginal necessity of relevance of a term t_i can be estimated by summing its weights in the documents while taking into account the number of documents in which it occurs. Therefore greater is the number of document where a term is relevant, greater is its marginal necessity of relevance. The proposed formula is the following:

$$\Pi(t_i) = 1 \ and \ N(t_i) = 1 - \Pi(\bar{t_i}) = \frac{\sum_{D_j} w_{ij}}{n_i} \quad (11)$$

with n_i being the number of document in the collection where the term t_i occurs,
w_{ij} is the weight of the term t_i in the document D_j.

2) Non-root term nodes

For each non-root term node T_i with parent set Par(T_i), we need to estimate the set of conditional possibility distributions, $\Pi(T_i|\theta^l)$, one for each possible configuration θ^l of parent nodes values. The proposed formulas are the following:

$$\Pi(t_i \left| \theta^l \right.) = \begin{cases} 1 if \ \exists d_k \in D \left| \forall t_j \in \theta^l, \theta_j^{d_k} = \theta_j^l \right. \\ 0 \quad otherwise \end{cases} \quad (12)$$

$$\Pi(\bar{t_i} \left| \theta^l \right.) = 1 - N(t_i \left| \theta^l \right.) = 1 - \frac{n(<t_i, \theta^l>)}{n(<t_i>) + n(<t_i, \theta^l>)} \quad (13)$$

With θ : is the set of possible configurations of the parent set of a term T_i,

$\theta^l \in \theta$: is one of the possible configurations of the parent set node of a term T_i,

$\theta_j^{d_k}$: is the instantiation of the term T_j in the document d_k,

θ_j^l : is the instantiation of the term T_j in θ^l ,

n(<...>): is defined as the number of documents containing all the terms that are included as relevant in the configuration θ^l and excluding those that are not relevant in it. In our model, this estimation is based on the Jaccard similarity measure which was also used for this task in [10], [12].

3) Document nodes

For each document node D_j, with a set of parents Par(D_j), we need to estimate a set of conditional possibility distributions $\Pi(D_j|\theta^l)$, one for each possible configuration θ^l of parent nodes values. Here, Par (D_j) is the set of term nodes used to index the document D_j. For instance, if a document is indexed by k terms, we need to estimate and store 2^k possibility values.

In order to reduce the estimation complexity which is due to the large number of terms by which a document is indexed, we propose to make these estimations using a Noisy OR-gate. It is generally used for this kind of problems [3], [23]. The proposed estimators are the following:

$$\Pi(d_j \left| \theta^l \right.) = \begin{cases} 1 \ if \ \exists \ T_i \in \theta^l \left| \theta_i^{D_j} = \theta_i^l = t_i \right. \\ 0 \quad otherwise \end{cases} \quad (14)$$

$$\Pi(\bar{d_j} \left| \theta^l \right.) = 1 - N(d_j \left| \theta^l \right.) = 1 - \frac{1 - \prod_{i \left| t_i = \theta_i^{D_j} \right.} q_{ij}}{1 - \prod_{t_k \in Par(D_j)} q_{kj}} \quad (15)$$

With θ : is the set of possible configurations of the parent set of the document D_j,

$\theta^l \in \theta$: is one of the possible configurations of the parent set of the document D_j,

$\theta_i^{D_j}$: is the instantiation of the term T_i in the document D_j,

θ_i^l : is the instantiation of the term T_i in θ^l ,

q_{ij}: is the weight w_{ij} of a term t_i in the document d_j. It is defined by the following formula which was proposed by Boughanem et a l[3]:

$$w_{ij} = \frac{\ln(\frac{N}{n_i})}{\ln(N)} \times \frac{tf_{ij}}{\max\limits_{\forall t_k \in d_j}(tf_{kj})} \quad (16)$$

with, N : is the number of document in the collection,
n_i : is the number of documents containing the term t_i, *and*
tf_{ij} : is the frequency of the term t_i in the document d_j.

D. The retrieval engine: inference in the PNIRM

Once the Possibilistic network is built, it can be used to retrieve documents that are relevant to a user query by means of the inference process. The model should be able to infer propositions like:

- ✓ It is is more or less plausible (to a certain degree) that the document is relevant for the user need, denoted by $\Pi(d_j|Q)$.
- ✓ It is almost certain (in possibilistic sense) that the document is relevant to the query, which is quantified by a degree of conditional necessity denoted by N(dj|Q).

A low value of $\Pi(d_j|Q)$ is meant to eliminate irrelevant documents (weak plausibility). If $\Pi(d_j | Q) = 0$, it is certain that document d_j is not relevant to the query Q. However $\Pi(d_j |$ Q)=1does not imply that the document is relevant, only that nothing prevents the document from being relevant. The second evaluation focuses attention on what looks very relevant.

Under a possibilistic approach, given a query *Q*, we are thus interested in retrieving necessarily relevant documents; or at least possibly relevant ones if there is none of the first kind. To achieve this task, the set of terms in the query will play the role of a new piece of evidence provided to the system. This information will be propagated toward the network nodes. Then, the documents will be sorted first by their posterior necessity of relevance $N(D_j|Q)$ and then by their posterior possibility of relevance $\Pi(D_j|Q)$ to the user query.

In order to do the inference task, we used the product-based possibilistic adaptation of the probabilistic propagation algorithm in junction trees proposed by Ben Amor [1].

V. Experimental evaluation and analysis of results

Several experiments have been conducted in order to evaluate the performance of the proposed model. We have applied our model to three well-known test document collections: CISI, MEDLARS and CRANFIELD. The main characteristics of these collections are shown in Table 1.

The choice of these document collections is justified by two facts:

1. Their medium size makes them the appropriate collections to evaluate our model preparing and tuning the latter to work with larger collections.
2. They have been used as test bed by the model to which we want compare our model. Thus, to give more rigors to results we have to use these collections in our experiments.

Table 1. Main features of the standard test collections

Collection	N. documents	N. terms	N. queries
CISI	1460	4985	112
CRANFIELDS	1398	3857	225
MEDLARS	1033	7170	30

For each one of these collections, several retrieval experiments were done in order to select the value of threshold (β) of terms dependency that yields optimal retrieval results. In fact, we have tested several values of this threshold, ranging from two up to ten. For each one of these values, a complete retrieval experiment has been conducted on each document collection. Though the best threshold value changes from a collection to another, we found that the value three yields a good performance on almost all the document collections and, therefore we decide to choose it for the rest of experiments.

In order to determine the effectiveness of the proposed model we have compared it to the Bayesian Network Retrieval Model (BNRM) [10] which breaks the independence assumption. The performance measure that we have used for the evaluation is the average precision for the eleven standard values of recall (AP-11). All the experimentations have been made using our own implimentation of the two models. They have been carried out on a single machine with Intel Core i3 2.20 GHZ CPU, 4 GB of RAM and 500 GB of local disk storage. The operation system used is Windows 7 Professional 64 bit edition. The C language was used both to build the structure and to estimate the parameters of the two models.

For the CISI and MEDLARS document collections, test results are presented in Figure 2 and Figure 3 respectively. It seems clear from the curves that our model performs better than the BNRM, which proves the efficiency of the proposed model.

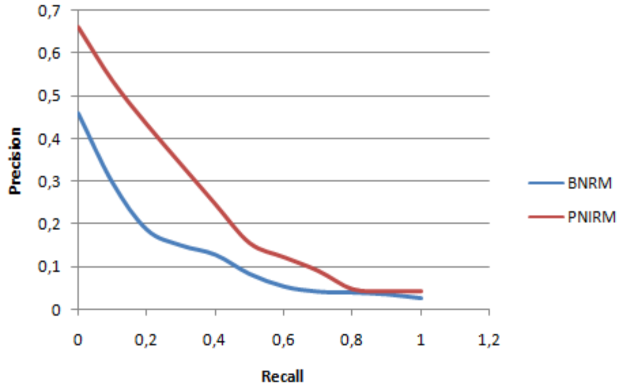

Figure 1. 11-point average precision curves for CISI collection.

The result for the CRANFIELDS collection is shown in Figure 4. It seems obvious from the curves that the BNRM model performs slightly better than the proposed model. More precisely, BNRM precision values are slightly better at recall levels ranging from 0.0 to 0.1. Then, the behavior of both models became almost the same for recall levels ranging from 0.1 to 0.3. Finally, our model performs slightly better for the rest of recall levels.

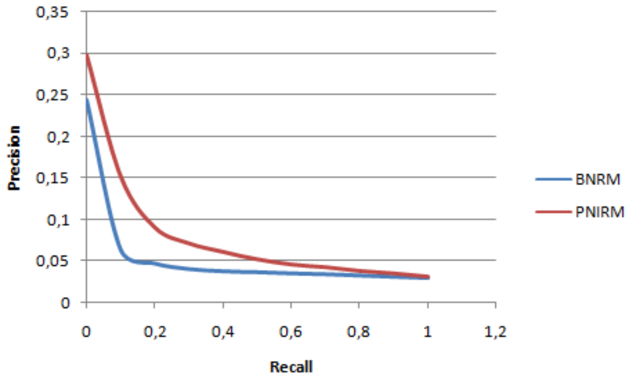

Figure 3. 11-point average precision curves for MEDLARS collection.

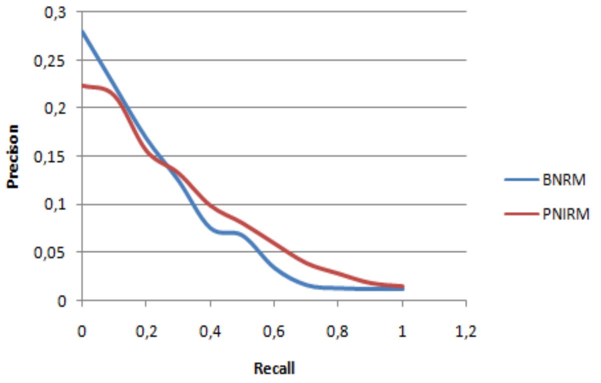

Figure 4. 11-point average precision curves for CRANFIELDS collection

VI. Conclusion and future work

In this paper, we have presented a new information retrieval model based on possibilistic networks that breaks the independence assumption between indexing terms. The main objective of this model was to focus only on the most important dependence relationships between terms. For that, we have developed a new approach that uses the strength of dependency between each pair of terms within each document as criterion for the identification of dependent pairs of terms. The quantification of the relevance of a document to a user query and the quantification of the strength of dependence relationships between pairs of terms are made using two possibilistic measures (i.e. the possibility and the necessity).

The performance of the proposed model was compared to the performance of an existing Bayesian network information retrieval model. Primary experimental results showed that it outperforms the other model on two medium-size standard document collections.

We propose as future research to evaluate the behaviour of the proposed model on big document collections with hundreds of thousands of terms such as the web in order to identify the amendments necessary to accommodate the latter to web context. Another line of research that we are considering is to develop a new mechanism to extract term dependence relationships based on semantic analysis of documents.

References

[1] N. Ben Amor, “Qualitative Possibilistic Graphical Models : From Independence to Propagation Algorithms”. PhD thesis, Universit é de Tunis, Institut Sup érieur de Gestion, Tunis, Tunisia, 2002.

[2] C. Bong-Hyun, L. Changki, G. L. Gary, “Exploring term dependences in probabilistic information retrieval model”, *Information Processing and Management,* vol. 39, pp. 505–519, 2003.

[3] M. Boughanem, A. Brini, D. Dubois, “Possibilistic networks for information retrieval”, *International Journal of Approximate Reasoning,* vol. 50, pp.957-968, 2009.

[4] J.P. Helen, W. Peter, “The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems ”, *Journal of the American Society for Information Science*, vol. 42, pp. 378-383, 1991.

[5] M.N. Omri, T Chenaina, “Uncertain and approximate knowledge representation to reasoning on classification with a fuzzy networks based system”. In *Proceedings of IEEE International Fuzzy Systems Conference Proceedings (FUZZ-IEEE) , pp. 1632-1637, 1999.*

[6] C. Borgelt, R. Kruse, “Evaluation Measures for Learning Probabilistic and Possibilistic Networks”. In *Proceedings of the 6th IEEE International Conference on Fuzzy Systems,* pp. 1034-1038, 1997.

[7] W. Chebil, L. F. Soualmia, M. N. Omri, S. J. Darmoni, “Indexing biomedical documents with a possibilistic network”, *Journal of the Association for Information Science Technology*, vol. 66, pp 227-231, 2015.

[8] W. Chebil, L. F. Soualmia, M. N. Omri, S. J. Darmoni, “Biomedical Concepts Extraction based on Possibilistic Network and Vector Space Model”. In *Proceedings of 15th Conference on Artificial Intelligence in Medicine (AIME'2015),* pp. 227-231, 2015.

[9] L. M. De Campos, “Independency relationships and learning algorithms for singly connected networks”, *Journal of Experimental & Theorical Artificial Intelligence*, vol. 10, pp. 511-549, 1998.

[10] L. M. De Campos, J.M. Fernandez-Luna, J. F. Huete, “The BNR model: foundations and performance of a Bayesian network-based retrieval model”, *International Journal of Approximate Reasoning,* vol (34), pp. 265-285, 2003.

[11] L.M., De Campos, J.M., Fernandez-Luna, J.F., Huete, “Clustering terms in the Bayesian network retrieval model: a new approach with two term-layers”, *Applied Soft Computing,* vol (4), pp. 149-158, 2004.

[12] S. Dongyu, Q. Zhengwei, F. Cheng, Y. Jinyuan, “An information retrieval model based on probabilistic network”. In *Proceedings of the IEEE International Conference on Services Computing (SCC 2004)*, pp. 423-426, 2004.

[13] D. Dubois, and H. Prade, Possibility Theory, Plenum Press, New York, 1988.

[14] K.Garrouch, M. N. Omri, B. Ayeb, “Pertinent Information retrieval based on Possibilistic Bayesian network : origin and possibilistic perspective”. In *Proceedings of the International Conference on Computing & e-Systems,* 2008.

[15] A. Elbahi, M.N. Omri, M.A.. Mahjoub, K. Garrouch, “Mouse Movement and Probabilistic Graphical Models Based E-Learning Activity Recognition Improvement Possibilistic Model”, *Arabian Journal for Science and Engineering*, pp. 1-16, 2016.

[16] K. Garrouch, M. N. Omri, A. Kouzana, “A New information retrieval model based on possibilistic Bayesian network”. In *Proceedings of the International Conference on Computer Related Knowledge (ICCRK'2012)*, pp. 79-88, 2012.

[17] K.Garrouch, M. N. Omri, “Possibilistic Network based Information Retrieval Model”, In *Proceedings of 15th International conference on Intelligent Systems Design and applications*, 2015.

[18] Y. Gupta, A. Saini , A.K. Saxena, “A new fuzzy logic based ranking function for efficient Information Retrieval system”, *Expert Systems with Applications,* vol (42), pp. 1223–1234, 2015.

[19] C. Keke, C. Chun, B. Jiajun, “Exploring of term relationship for Bayesian network based sentence retrieval”, *Pattern Recognition Letter*, vol (30), pp. 805-811, 2009.

[20] H.S. Kim, I. Choi, M. Kim. “Refining term weights of Documents using term Dependencies”. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 552-553, 2004.

[21] F. Naouar, L. Hlaoua, M.N. Omri, “Possibilistic Information Retrieval Model Based on Relevant Annotations and Expanded Classification”. In *Proceedings of the 22nd International Conference on Neural Information Processing(ICONIP2015)*, pp. 185-198, 2015.

[22] M. N. Omri, “Fuzzy knowledge representation, learning and optimization with Bayesian analysis in

fuzzy semantic networks". In *Proceedings of the 6th International Conference on Neural Information Processing (ICONIP'99*), pp. 412-417, 1999.

[23] S. Parsons, J. Bigham, "Possibility theory and the generalized noisy or model". In *Proceedings of the 6th International Conference on Information Processing and Management of Uncertainty*, pp. 853-558, 1996.

[24] G. Rebane, J. Pearl, "The recovery of causal polytree from statistical data", *Uncertainty in Artificial Intelligence*, vol (3), pp.175-182, 1989.

[25] M. Robert, Jr. Losee, "Term dependence truncating the Bahadur Lazarsfeld expansion", *Information Processing & Management*, vol (30), pp. 293–303, 1994.

[26] G. Salton, C. Buckley, "Term weighting approaches in automatic text retrieval", *Information Processing & Management*, vol (24), pp.513-523, 1998.

[27] H. Turtle, B. Croft. "Inference networks for document retrieval". In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 1–24, 1990.

[28] J. M. Xu, W.S. Tang, J.M. Xu, Z.Y. Chen, Z.H. Luo, "A Word Similarity Based Belief Network IR Model with Two Term Layers". In *Proceedings of 1st WRI Global Congress on Intelligent Systems (GCIS '09)*, pp. 514-517, 2009.

[29] L.A. Zadeh, "Fuzzy sets as a basis for a theory of possibili*ty", Fuzzy Sets and Systems*, vol (1), pp.9–34, 1978.

Author Biographies

Kamel Garrouch is currently a PHD student at Department of Computer Science, Faculty of Science of Monastir, University of Monastir Tunisia. He received his Master degree in Computer Science (Software Engineering) from Tunis University Tunisia in 2006. He published several papers on information retrieval using possibilistic networks. His research interest includes information retrieval, possibilistic networks and Bayesian networks.

.

Mohamed Nazih OMRI received his Ph.D. in Computer Science from Jussieu University, in 1994. He is a Professor in computer science at Monastir University. From January 2011, he served as the Director of MARS (Modeling of Automated Reasoning Systems) Research Unit.

His group conducts research on Approximate reasoning, Fuzzy logic, Modeling of complex systems, web information retrieval, Bayesian and Semantic Networks.

.