Information Retrieval Model Based on Possibilistic Ontology

Rihab Ben Lamine, Mohamed Nazih Omri
Mars Unit of Research
Department of Computer Sciences
Faculty of Sciences of Monastir
Monastir, Tunisia
benlaminerihab@yahoo.fr, MohamedNazih.omri@fsm.rnu.tn

ABSTRACT: *Classical approaches of Information Retrieval treat documents and queriesas «bags of words» with no syntax and no semantic.This representation is based on the co-occurrence of words in a textand does not take into account the semantic relationsth at exist between them. With the advent of ontologies,at the end of the 1990s, new approaches of Information Retrieval have appeared. These approaches attempt to capture the semantics of documents and queries by representing them with ontology concepts.*

A classical ontology doesnot avoid the vagueness and uncertainty that characterize human reasoning, in modeling the real world.Using the theory of possibility seems to be a solution. In this paper, we propose an indexing model based on possibilistic ontology. Therefore, we present a method for converting a conventional ontology into a possibilistic one and then we defineasemantic similarity measuresuitable for possibilistic ontology. The weights of index terms are based on the use of this measure.

Keywords: Information Retrieval, Ontology, Possibility Theory, Semantic Indexing, Conceptual Indexing

Received: 25 December 2012, Revised 29 January 2012, Accepted 3 February 2012

1. Introduction

With the advent of ontologies, new information Retrieval approaches are developed. These approaches try to capture the semantics of documents and queries by representing them with concepts of ontologies [1].

Ontology is a knowledge modeling of a given domain.

According to [2], our knowledge is often imperfect, either because they are uncertain (there is a doubt on their validity) or vague (there is a difficulty in expressing them clearly).

So, the real world appears at the same time vague and uncertain and the borders of the states of nature are not always very clear.

Since a classical ontology does not avoid the imperfection and the imprecision of the real world, using possibilistic ontology seems to be a solution. In fact, in such ontology relations between concepts are characterized by possibility degree that quantifies the possibility of this relationship, and necessity degree that quantifies the certainty of a relationship between two concepts.

2. Related Works

In the context of using possibilistic ontology in Information Retrieval, we can mention the work of Loiseau [3]. In his approach of Information Retrieval, an adaptation of fuzzy pattern matching to linguistic terms is proposed. It is based on the idea to retrieve information that may not match exactly the query. Therefore, a possibilistic ontology, where the relations between terms are modeled by the possibility and the certainty that their meanings refer to the same thing, following the idea suggested in [4], is used. The similarity between terms is independent of the hierarchical distance between terms in the ontology. For two labels t_i and t_j, used to describe a piece of information:

$\Pi (t_i , t_j) = \Pi (t_i , t_j)$ represents to what extent t_i and t_j can describe the same thing.

$N (t_i , t_j)$ represents to what extent it is certain that t_j is a specialization of t_i . If is a perfect specialization of t_i, then $N (t_i, t_j) = 1$.

The following properties must be satisfied by these measures:

- Reflexivity: $\Pi (t_i , t_j)= 1$.
- Symmetry: $\Pi (t_i , t_j) = \Pi (t_i , t_j)$.
- $\Pi (t_i , t_j) \geq N (t_i , t_j)$, since specialization entails that the meanings overlap.
- $N (t_i , t_j) > 0 \Rightarrow \Pi (t_i , t_j)$. If t_i specializes certainly t_j, then it must be fully possible that they refer to the same thing.

The degrees presented in the ontology are defined between every pair of concept linked by a relation in the ontology. They can be completed using previous properties and the two following forms of transitivity:

- The specialization transitivity [5]:

$$N (t_i , t_j) \geq \min (N(t_i , t_k), N (t_k , t_j))$$

- The "*hybrid transitivity*":

$$\Pi (t_i , t_j) \geq N (t_i , t_j) * \Pi (t_k , t_j)$$

where

$$a * b = b \text{ if } b > 1 - a, \quad 0 \text{ otherwise.}$$

In qualitative pattern matching, a set of terms T_i is associated to each information domain i and these terms are related throughout possibilistic ontology O_i.

In this information retrieval model, a textual data is represented as a set of words. Only significant words are retained. A key feature of this model is indexing documents within the ontology.

To be homogenous with the ontology representation, the relevance between a query and a document and the correspondence between concepts and document must be estimated by possibility and necessity degrees, taking into account the weight of document terms. The weight of a term t_i in a document D_j is estimated by combining the frequency tf_{ij} of t_i in D_j and the inverse frequency $idf_i = log (d / df_i)$, where df_i is the number of documents containing the term t_i and d is the total numberof documents.

The weight ρ_i^k of a term t_i in a document D_j is considered as an intermediary degree between possibility and necessity that this term describes D_j. The possibility and the necessity are computed as follow [6]:

$$\Pi (t_i , D_j) = 2 * \rho^j_i \ ; N(t_i , D_j) = 0 \ if \ \rho^j_i < 0.5 \qquad (1)$$

$$\Pi (t_i , D_j) = 1 \ ; N(t_i , D_j) = 2 * \rho^j_i - 1 \ if \ \rho^j_i \ otherwise$$

By indexing documents within the ontology, queries can be evaluated according to qualitative pattern matching properties. Thus, the relevance of a document for a query is computed by a couple of possibility and necessity degrees:

$$rsv (R , D) = (\Pi (R , D), N (R, D)) \qquad (2)$$

These two values for a document *D* and a query Q are calculated using (1) and:

$$\Pi (R, D) = min_k \, max \, (1 - w_k , max_{i, j} \, min \, (\lambda_k^j , \Pi (t_k^i , t_i) , Đ (t_i , D)) \qquad (3)$$

$$N (R, D) = min_k \, max \, (1 - w_k , max_{i, j} \, min \, (\lambda_k^j , N (t_k^i , t_i) , N (t_i , D)) \qquad (4)$$

where

$$D = \{t_i , t_\mathrm{i} \in T \},$$

R = $\Lambda_\mathrm{k} (w_k , \mathrm{V}_j (\lambda_k^j , t_k^i))$ with $t_k^i \in T$, *T* is the global domain and $\lambda_k^j , w_k \in [0, 1]$ et $max_k w_k = max_j \lambda_k^j = 1$.

The weight λ_k^j reflects how satisfactory this term is for the user and the weight w_k expresses how compulsory is each elementary requirement in the conjunction forming the query.

This approach was tested on a small collection containing about 200 titles of computer sciences articles, and a simple ontology corresponding to these titles terms is used. Thus, the obtained results are not decisive as for the evaluation of real-size system, given the small size of the collection and small number of queries.

3. Proposed Model

Our approach is related to conceptual indexing domain. In a document, there are important terms (informative), necessary to represent it, therefore necessary to decide on the relevance of the document to a query, and other less informative terms which are possibly interesting for representing the document. Our approach differs from Loiseau ‘s approach in the fact that in our approach we proposed to keep the attitude commonly used in conceptual and semantic indexing which is based on similarity measures known in literature. For that, we propose to adapt a similarity measure to possibilistic ontology.

The first step of any Information Retrieval System is indexing. In our model, we propose an indexing method based on possibilistic ontology. After projecting the corpus on the ontology, each document is represented by a set of concepts. For each concept of a document we propose to calculate a global weight which is composed of two parts: a semantic weight which the calculation is based on semantic similarity measure which we propose to adapt to possibilistic ontology, and a statistical weight which the calculation is based on statistical measures such as the frequency of concept occurrence in a document and the frequency of concept occurrence in the corpus.

3.1 Possibilistic Ontology

In our model, possibilistic ontology is an ontology where relations between concepts are modeled by necessity and possibility degrees.

Π (*C*, *Parent*(*C*)) represents to what extent relation between *C* and *Parent* (*C*) is possible.

N (*C*, *Parent*(*C*)) represents to what extent relation between *C* and *Parent* (*C*) is certain.

We choose to use the MeSH[1] thesaurus which is a medical thesaurus used to index the MEDLINE bibliographic database. It provides a hierarchical and associative organization and includes up to eleven levels deep.

In order to transform this ontology to a possibilistic one we propose an algorithm to attribute possibility and necessity degrees to each relation in the ontology. The different steps of this algorithm are as follow:

- First step: possibility degrees attribution:

As starting point is an already existent ontology, weassume that if a relation between two concepts exists in this initial ontology, then this relation is totally possible.

So, for each pair of concept (*C*, *Parent* (*C*)), we have:

$$\Pi (C, Parent\ (C)) = 1 \qquad (5)$$

- Second step: necessity degrees attribution:

[1] http://www.nlm.nih.gov/mesh/MBrowser.html

A MeSH concept can belong to more than one hierarchy, and every time a different parent. Because of this, a concept C_i having more relations with different parents than another concept C_j, must have necessity degrees with its parents that are lower than necessity degrees that model relations between C_j and its parents.

We can remark that the necessity degree of relation between a concept *C* and its parent *Parent* (*C*) depends on the number of distinct parents of the concept *C*. So, necessity degrees can be computed as follow:

$$N\,(C, Parent\,(C\,)) = 1 - ((\,nbreParents\,(C\,) - 1) \,/\, nbreParents\,(C\,)) \tag{6}$$

where *nbreParents*(*C*) is the number of distinct parents of *C* and (*nbreParents*(*C*) – 1) / *nbreParents* (*C*)) represents the participation of the concept *C* in the senses of its other parents which are different from *Parent* (*C*).

Necessity degree between a concept and its parent is maximal if this concept has no other relation with other concepts, and more a concept has other relations, more the necessity degree between it and its parent will be lower.

3.2 Adaptation of semantic similarity measure to possibilistic ontology

Many semantic similarity measures have been proposed (about twelve) using semantic or hierarchical network structures. We choose Wu-Palmer measure [7] because it is simple to implement and offered good performances [8]. The expression of this measure is given by the following formula:

$$sim_{W\&P}\,(C_i\,, C_j\,) = 2 * prof\,(C\,) \,/\, (dist\,(C_i\,, C\,) + dist(C_j\,, C\,) + 2 * prof\,(C\,)) \tag{7}$$

where *C* is the most specific common concept, *dist* (C_i , *C*), is the distance between concept C_i and *C* and *prof* (*C*) is the distance between the ontology root and the most specific common concept *C*. The distance between two concepts is measured by the number of links from the first concept to the second one. We can remark that the distance between two concepts does not consider the uncertainty of the relation between them. Indeed, the distance is always equal to 1 between a concept and its parent. Because of this, we propose to integrate uncertainty by using ontology's necessity and possibility degrees into the computation of distances between concepts. We define possible distance, which represents the distance that possibly separates two concepts and necessary distance, which necessarily separates two concepts. Possible distance between concept C_i and concept C_k can be computed as follow:

$$dist_{\Pi}\,(C_i\,,C_k\,) = \sum_{j\,=\,i}^{k\,-\,1} (1 \,/\, \Pi\,(C_j\,, C_{j\,+1}\,)) \tag{8}$$

where C_{j+1} is the parent of C_j.

This possible distance considers the possibility that characterizes the semantic relation between C_i and C_k. The necessary distance between concept C_i and concept C_k can be computed by:

$$dist_N\,(C_i\,,C_k\,) = \sum_{j\,=\,i}^{k\,-\,1} (1/N\,(C_j\,, C_{j\,+1}\,)) \tag{9}$$

where C_{j+1} is the parent of C_j.

This necessary distance considers the certainty that characterizes the semantic relation between C_i and C_k .

Considering these two possibilistic distances, Wu-Palmer measure is transformed into two similarity measures: possible similarity measure which represents to what extent two concepts are possibly similar, and necessary similarity measure which represents to what extent two concepts are certainly similar. The possible similarity measure between two concepts C_i and C_j is given by:

$$Sim_{\,\Pi}\,(C_i\,, C_j\,) = 2 * dist_{\,\Pi}\,(root\,, C\,) \,/\, dist_{\,\Pi}\,(C_i\,, C\,) + dist_{\,\Pi}\,(C_j\,, C\,) + 2 * dist_{\,\Pi}\,(root\,, C\,) \tag{10}$$

and the necessary similarity between two concepts C_i and C_k is given by:

$$Sim_N\,(C_i\,, C_j) = 2 * dist_{\,N}\,(root\,, C\,) \,/\, dist_{\,N}\,(C_i\,, C\,) + dist_{\,N}\,(C_j\,, C\,) + 2 * dist_{\,N}\,(root\,, C\,) \tag{11}$$

3.3 Terms weighting

According to Desmontils and Jacquin [9], the representative power of a concept takes into account not only the occurrence frequency of terms that designate the concept in documents but also its relations with other concepts of document. The more a concept has relations with other concepts of a document, the more it is representative of the document.

In our model, for each concept in a document is given a semantic weight and statistical weight.

3.3.1 Semantic weight

Semantic weight computation is based on possibilistic semantic similarity measures given by (10) and (11). For each concept C_i of document D_k is given possible semantic weight which reflects to what measure C_i can possibly describe D_k , and necessary semantic weight which reflects to what measure C_i can certainly describe D_k.

The possible semantic weight of a concept C_i in a document D_k , which is the average of possible semantic similarity between the concept C_i and others concepts in the document D_k , can be calculated by:

$$p_{poss}(C_i , D_k) = \sum_{\forall C_j \in D_k , C_j \neq C_i} Sim_{\Pi}(C_i , C_j) / m \qquad (12)$$

where m is the number of concepts of the document D_k that are different from C_i. The same for the necessary semantic weight, it can be calculated for a concept C_i and a document D_k by:

$$p_{necess}(C_i , D_k) = \sum_{\forall C_j \in D_k , C_j \neq C_i} Sim_{N}(C_i , C_j) / m \qquad (13)$$

3.3.2 Statistic weight

As in [3], we consider the weight [3] of a term t_i in a document D_k as an intermediate degree between possibility and necessity that this term describes D_k. So, the weight ρ_i^k is given by:

$$\rho_i^k = tf_{ik} * idf_i \qquad (14)$$

Where $tf_{ik} = freq_{ik} / max_j\ freq_{jk}$, $freq_{ik}$ is t_i frequency in the document D_k, and $max_j\ freq_{jk}$ is the frequency of the most frequent term in the document D_k. idf_i is the inverse of document frequency which represents the term importance in the collection of documents and it is given by:

$$idf_i = log\ (\mathrm{N} / n_i)$$

Where N is the total number of documents and n_i is the number of documents where the term t_i appears. To be in accordance with possibilistic theory, the weight ρ_i^k must be normalized. So:

$$\rho_i^k = \rho_i^k / max_j\ \rho_j^k$$

$max_j\ \rho_j^k$ is the maximal weight in the document D_k.

The possible statistic weight and the necessary statistic weight of a concept C_i in a document D_k , $\Pi (C_i , D_k)$ and $N (C_i , D_k)$ are calculated as follow:

$$\begin{aligned} &\Pi\ (C_i , D_k) = 2 * \rho_i^k ;\ N(C_i , D_k) = 0\ if\ \rho_i^k < 0.5\ and \\ &\Pi\ (C_i , D_k) = 1;\ N(C_i , D_k) = 2 * \rho_i^k \text{ -1 } otherwise \end{aligned} \qquad (15)$$

3.3.3 Global weight

As in [9], we consider that global weight of a concept C_i in a document D_k ($g\_w\ (C_i , D_k)$) is a linear combination of the semantic weight ($sem\_w\ (C_i , D_k)$) and the statistic weight ($stat\_w(C_i , D_k)$) (see (16)).

$$g\ \_w\ (C_i , D_k) = (\alpha * sem\_w\ (C_i , D_k)) + \beta * stat\_w\ (C_i , D_k) / (\alpha + \beta) \qquad (16)$$

Then, each concept has a possible global weight which is a linear combination of its possible semantic weight and its possible statistic weight, and a necessary global weight which is a linear combination of its necessary semantic weight and its necessary statistic weight. We choose $\alpha = 1$ and $\beta = 1$, because we want that semantic weight and statistic weight intervene in the same way in calculating the global weight.The possible global weight is calculated for a concept C_i in a document D_k by:

$$w_{poss}(C_i, D_k) = (\Pi(C_i, D_k) + p_{poss}(C_i, D_k)) / 2 \qquad (17)$$

and the necessary global weight for a concept C_i in a document D_k is calculated by:

$$w_{necess}(C_i, D_k) = (N(C_i, D_k) + p_{necess}(C_i, D_k)) / 2 \qquad (18)$$

3.4 Query Evaluation

For a document D and a query Q, a possible relevance and a necessary relevance are calculated.

We wish to exploit semantic similarity between document and query concepts. For that, we propose to use the relevance measure proposed in [10] and expressed as follow:

$$sim\,(Q, D) = \sum_k \sum_l w_k * w_l * sim\,(C_k, C_l) \,/ \sum_k \sum_l w_k * w_l \qquad (19)$$

where w_k is the weight of a concept C_k in the document D and w_l is the weight of the concept C_l of the query Q.

Then, this relevance measure is transformed into two relevance measures: possible relevance given by:

$$Poss\_rel\,(Q, D) = \sum_k \sum_l (w_{k\,,\,poss} * w_l * sim_{\Pi}\,(C_k, C_l) \,/ \sum_k \sum_l w_{k\,,\,poss} * w_l \qquad (20)$$

and necessary relevance:

$$Ness\_rel\,(Q, D) = \sum_k \sum_l (w_{k\,,\,necess} * w_l * sim_N\,(C_k, C_l) \,/ \sum_k \sum_l w_{k\,,\,necess} * w_l \qquad (21)$$

Documents are then ranked in decreasing order of necessary relevance and in the case of equality in decreasing order of possible relevance.

4. Experiments

In order to evaluate the performance of our information retrieval model, we used the MuchMore[2] [11] corpus which contains 7823 documents, 25 queries and relevance judgments. For queries, only concept detection is applied. The documents of the collection are summaries of Springer articles and treat medical domain. The language used is largely covered by MeSH. Our system, named RIOP, is then compared to Loiseau's system which is applied on the MuchMore corpus using possibilistic MeSH ontology.

The obtained results allow us to dress precision and recall comparison diagrams shown in Figure 1 and Figure 2 respectively, and Averaged 11-point precision/recall diagram shown in Figure 3.

Observing Figure 2, we can see that our model RIOP is more performant in terms of precision. According to Figure 3, at the beginning of retrieved documents list RIOP model gives a better precision than the other model, expect, at the end of the list, the two models have almost the same performance. This equality is not of a big influence because in practice an internet user rarely explores the whole list of retrieved documents. Finally, we can see the positive impact of using our possibilistic similarity measures in terms weighting.

From Figure 1, we note that Loiseau's system is more performant in terms of recall. This can be due to our query evaluation method. For that we tried to use query evaluation presented by Loiseau, which we combined with our indexing method. This new model is named RIOP-Loiseau since it is a hybrid model between RIOP model and Loiseau's model.

The obtained results allow us to dress precision and recall comparison diagrams shown in Figure 4 and Figure 5 respectively, and Averaged 11-point precision/recall diagram shown in Figure 6.

[2] http://muchmore.dfki.de/

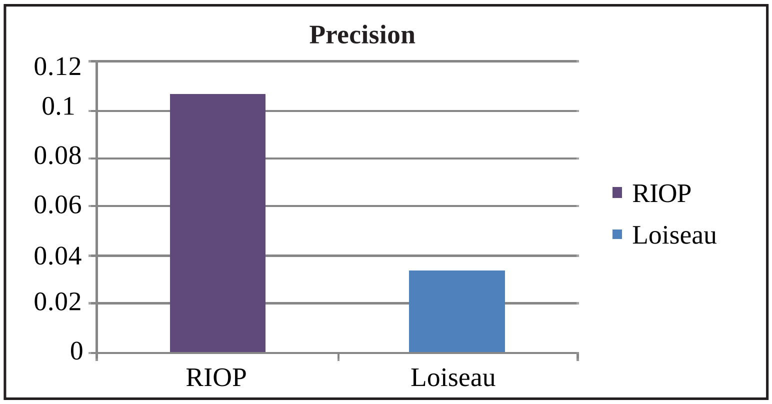

Figure 1. Precision rate comparison between RIOP and Loiseau'system

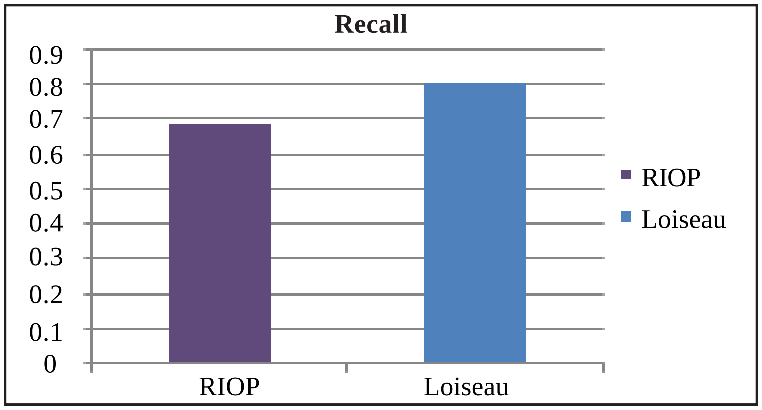

Figure 2. Recall rate comparison between RIOP and Loiseau's model

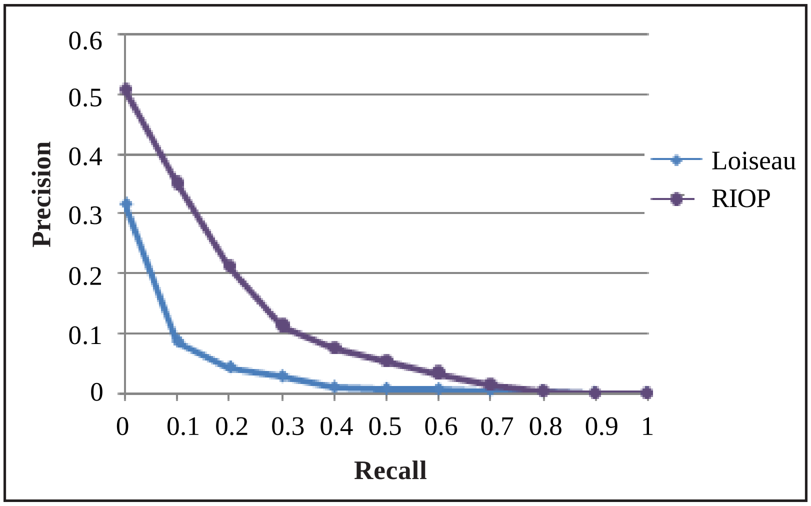

Figure 3. Averaged 11-point precision/recall graph of RIOP model and Loiseau's model

From Figure 4 and Figure 6, we see that our model RIOP is more performant than the RIOP-Loiseau model and the model of Loiseau. We can see also, the positive impact of integrating semantics in terms weighting, comparing the RIOP-Loiseau model and the model of Loiseau in Figure 6. However, as shown in Figure 5, Loiseau's model is still more performant in terms of recall.

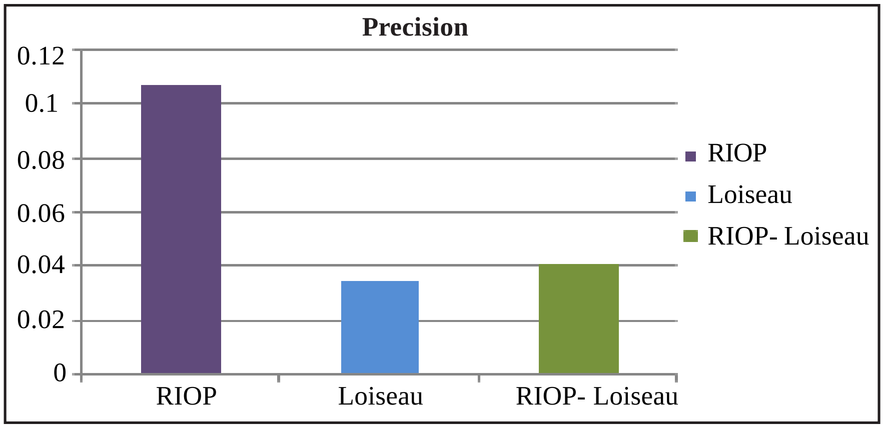

Figure 4. Precision rate comparison between RIOP, Loiseau's model and RIOP-Loiseau model

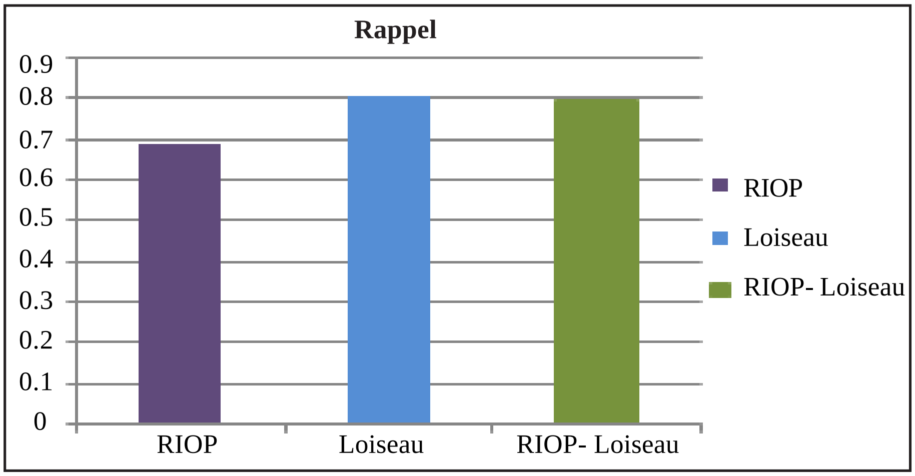

Figure 5. Recall rate comparison between RIOP, Loiseau's model and RIOP-Loiseau model

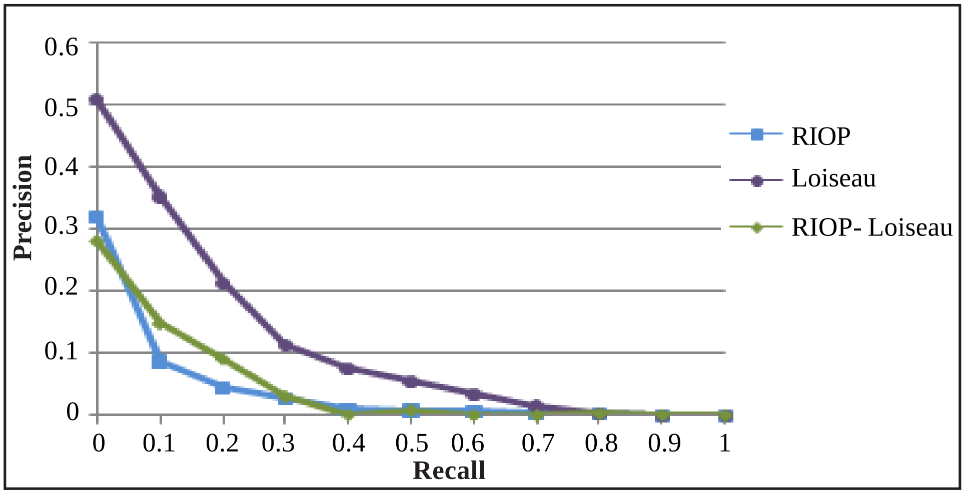

Figure 6. Averaged 11-point of precision/recall graphof
RIOP model, Loiseau's model and RIOP-Loiseau model

5. Conclusion

This paper presents a new IR approach based on possibilistic ontology. The originality of this approach is integrating semantics and uncertainty in terms weighting and relevance measure. The results show that our model, RIOP, is promising as we seem to obtain better average precision than the model proposed by Loiseau and our hybrid model RIOP-Loiseau.

References

[1] Baziz, M. (2005). Indexation conceptuelle guidée par ontologie pour la recherche d'information. Ph.D. thesis, Université de Paul Sabatier, Toulouse.

[2] Bouchon-Meunier, B. (1994). La logique floue, PUF collection Que sais-je ?, p. 2702.

[3] Loiseau, Y. (2004). Recherche flexible d'information par filtrage flou qualitatif, Ph.D. thesis, Université Paul Sabatier, Toulouse.

[4] Farreny, H., Prade, H. (1986). Dealing with vagueness of natural languages in man-machine communication. *In:* Karwowski, W. et Mital, A., éditeurs: Applications of Fuzzy Set Theory in Human Factors,Elsevier, p. 71-85.

[5] Rossazza, J., Dubois, D., Prade, H. (1997). A hierarchical model of fuzzy classes, *In*: Fuzzy and Uncertain Object-Oriented Databases, R. de Caluwe (Eds.), World Pub. Co, Singapore, p. 21-62.

[6] Prade, H., Testemale, C. (1987). Application of possibility and necessity measures to documentary information retrieval, LNCS, 286, p. 265-275.

[7] Wu, Z., Palmer, M. (1994). Verb semantics and lexical selection, *In*: Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, p. 133-138.

[8] Nguyen, B., Varlamis, I., Halkidi, M., Vazirgiannis, M. (2003). Construction de classes de documents web, Premières Journées Francophones de la toile, p. 173-182.

[9] Desmontils, E., Jaquin, C. (2002). Indexing a web site with a terminology oriented ontology, The Emerging Semantic Web, IOS Press, ISBN 1-58603-255-0, p. 181-197.

[10] Ioannis, V. (2005). Semantic similarity methods in wordNet and their application to information retrieval on the web, *In*: Proceedings of the 7th annual ACM international workshop on web information and data management, p. 10-16.

[11] Buitelaar, P. et al. (2004). Evaluation Resources for Concept-based Cross-Lingual IR in the Medical Domain. *In*: Proc. of LREC2004, Lissabon, Portugal, May.