
Evaluation of Term-based Queries using Possibilistic Ontologies

Yannick Loiseau, Mohand Boughanem, and Henri Prade

Institut de Recherche en Informatique de Toulouse,
118 route de Narbonne,
31062 Toulouse cedex 4
{loiseau,bougha,prade}@irit.fr

Summary. In multiple-source information systems, attribute values are often assessed in linguistic terms belonging to different vocabularies. The request itself, which may include preferences, may be expressed using terms of another vocabulary, raising the problem of matching the query and the information in a semantic manner. The fuzzy pattern matching framework allows us to compute matching degrees between queries and data represented by fuzzy sets, even if they do not perfectly match. The qualitative pattern matching no longer requires a fuzzy set representation thanks to the use of ontologies for computing similarity degrees between terms. This allows us to deal with information querying in face of heterogeneous sources of information. This chapter presents this tool and its application to database and textual information retrieval on two examples.

1 Introduction

In information retrieval (IR) as well as in databases systems, the vocabularies used to state the linguistic information are often heterogeneous when documents come from different sources. The terms used to express categories associated to a concept by one of the sources do not always correspond to the terms used by the other sources. For instance, documents written by different authors, from different journals, even if they deal with the same field, may use slightly different vocabularies. Moreover, even in the case of a unique source, or if the different sources use the same vocabulary, queries specifying user needs may be stated in another vocabulary, since the user does not know a priori which terms are used in the documents collection. This issue is still more obvious when using documents written in different languages, as it is the case in multilingual information retrieval [1]. This raises the problem of putting into correspondence the terms used in the different vocabularies. Moreover, the user may want to express preferences, under the form of flexible queries (e.g. [2-4]), about her/his requirements that have to be taken into account in the query evaluation process.

The fuzzy pattern matching framework [5] provides a tool for evaluating flexible queries in face of possibly imprecise data, where each linguistic label, both in the data or in the query, is represented by a fuzzy set. When these fuzzy sets are defined on the same domain, they can be compared using two indices called possibility and necessity measures. However, fuzzy pattern matching requires fuzzy set-based representations, usually associated with terms on numerical domains. In order to deal with more general linguistic terms, i.e. evaluating queries on documents where both of them are described by linguistic keywords that cannot be easily represented by fuzzy sets on a domain, the fuzzy pattern matching idea has been adapted to symbolic labels [6, 7] recently. Then, the matching between query terms and data no longer require the identity of terms, but rather a *semantic similarity* is computed between these terms, in a qualitative matching process. The relations between terms are now provided by means of an ontology estimating approximate synonymy and hypernymy relations between terms by means of possibility and necessity degrees. This *possibilistic ontology* is then used to evaluate relations that are not directly specified between terms of the ontology. This leads to an implicit query expansion. Thus it is possible to match a query and a piece of information taking into account preferences, even though terms do not match perfectly. This can partially solve the heterogeneous vocabularies issue.

The approach may appear to be reminiscent of thesaurus or terms association-based IR (e.g. [8]), where co-occurrence degrees are attached to pairs of terms. However, fuzzy pattern matching uses two degrees that respectively assess to what degree terms are possibly synonymous, or to what degree one is a specialization of the other. These two estimates differ from thesauri degrees. Indeed, terms such as *database* and *query* may be found co-occurrent, while *query* and *request* are not, even though the two latter terms are synonymous, while the two former are not.

The key notions of the fuzzy pattern matching used as a conceptual basis for the approach are first recalled. Its qualitative counterpart is then presented, where the semantic proximity of terms is assessed by means of possibility and necessity relations and propagated within possibilistic ontologies. Related works on weighted ontologies are also briefly surveyed. This chapter illustrates the use of this method on practical information sources, such as a tourism database and a collection of document titles. Finally, an extension of this model to full-text IR is outlined.

2 From Fuzzy to Qualitative Pattern Matching

2.1 Fuzzy Pattern Matching

The fuzzy pattern matching method has been developed in the fuzzy sets framework [5, 9, 10]. It is used to formulate flexible queries using fuzzy sets, evaluated on imprecise or fuzzy data also represented by fuzzy sets. This

technique estimates to what extent it is possible and to what extent it is certain that data represented by imprecise attributes fulfill a flexible query representing the user needs and preferences.

A pattern is a set of elementary requirements represented by labels encoding properties defined on the attributes domains. As an example, the pattern “cheap and large” associated with a database of houses, states that the “price” and the “size” attribute values of the searched house should be respectively compatible with *cheap* and *large*. Each label such as *cheap* and *large* is associated to a fuzzy set membership function which restricts values that are more or less compatible with the label meaning. These values belong to the domains of the corresponding attributes. Here, *cheap* and *large* are associated with membership functions defined on the price and size domains, which can be numerical ranges or discrete sets of values. Moreover, data are also described by sets of labels associated to fuzzy sets. These sets are possibility distributions representing imprecise or fuzzy data.

Let R and D be a pattern label and a piece of data respectively, that must be compared. They belong to the same domain U . Let μ_R be the membership function associated to the label R and π_D be the possibility distribution corresponding to D , which are functions from U to $[0, 1]$. Let u be an element of U . $\mu_R(u)$ is the compatibility degree between the value u and the meaning of R . $\mu_R(u) = 1$ means a complete compatibility with R and $\mu_R(u) = 0$ means a complete incompatibility with R . In the same way, $\pi_D(u)$ estimates the possibility that u is the value of the considered attribute describing the data. D is a fuzzy set of *possible* values, among which only one is the true value of the uncertain piece of data, whereas R is a fuzzy set of *more or less* compatible values. More precisely, $\pi_D(u) = 1$ means that u is totally possible, whereas $\pi_D(u) = 0$ means that u is not at all a possible value for the element. However, distinct values u and u' such that $\pi_D(u) = \pi_D(u') = 1$ may exist. In the following, μ_R and π_D are normalized, that is $\max_U \mu_R(u) = 1$ and $\max_U \pi_D(u) = 1$.

Two measures are defined to estimate the compatibility between a request element R and its counterpart D in the data: a possibility degree $\Pi(R; D)$ and a necessity degree $N(R; D)$ defined as [9]:

$$\Pi(R; D) = \sup_{u \in U} \min(\mu_R(u), \pi_D(u)) , \quad (1)$$

$$\begin{aligned} N(R; D) &= 1 - \Pi(\bar{R}; D) \\ &= \inf_{u \in U} \max(\mu_R(u), 1 - \pi_D(u)) . \end{aligned} \quad (2)$$

The possibility measure $\Pi(R; D)$ estimates to what extent it is possible that R and D refer to the same value u . It represents the *intersection* of the fuzzy set of values compatible with R with the fuzzy set of possible values for D . The necessity $N(R; D)$ measures to what extent it is certain that the value corresponding to D is compatible with R . It is an *inclusion* degree of the possible values for D into the set R of values compatible with the query label.

This measure $N(R; D)$ reflects the asymmetry between the user needs and the information itself.

It can be shown that $\Pi(R; D) \geq N(R; D)$. Note that if D is precise, that is: $\exists t', \pi_D(t') = 1$ and $\forall u \neq t', \pi_D(u) = 0$, i.e. $D = \{t'\}$, then $\Pi(R; \{t'\}) = N(R; \{t'\}) = \mu_R(t')$. Note also that if $\mu_R = \pi_D$, $\Pi(R; D) = 1$, and if R is a fuzzy set, $1 \geq N(R; D) \geq \frac{1}{2}$. In the case of continuous membership functions on real domains, (2) gives $N(R; D) = \frac{1}{2}$ if $\mu_R = \pi_D$. This reflects that there are values that are possible at degree 0.5 that are not in the 0.5 level-cut of R , i.e. in $\{u, \mu_R(u) \geq \frac{1}{2}\}$.

Classical fuzzy pattern matching recalled here is used in the following as a basis for defining *qualitative pattern matching*. The idea of qualitative pattern matching is to state fuzzy semantic relations between linguistic terms, using possibility and necessity degrees, now obtained from *possibilistic ontologies*.

2.2 Possibilistic Ontology

Fuzzy sets can interface numerical values with linguistic terms, using membership functions, and the comparison of terms can be evaluated by fuzzy pattern matching. However, the terms used for describing information need to be represented by a fuzzy set on a clearly identified domain, which is a severe limitation for information retrieval purposes. Similarity measures between words have been extensively studied in information retrieval literature, using for instance distance between nodes in a taxonomy, or based on a common information probability (e.g. [11]). Besides, a usual method when a query fails is to replace it by similar queries generated using ontologies or thesauri (e.g. [12]). Qualitative pattern matching remedies limitations of fuzzy pattern matching by enlarging the meaning of a term through the fuzzy set of its similar terms.

In *qualitative pattern matching*, a set of terms \mathcal{T}_i is assumed to be associated to each information domain i . \mathcal{T}_i is the set of labels that can be used to describe information in domain i . More precisely, $\mathcal{T}_i = \{t_i^j, j = 1, n(i)\}$, where t_i^j is a label (e.g. *hotel*), that can be used in order to describe a piece of information (here, an accommodation place). Contrary to standard fuzzy pattern matching, terms, or more generally symbolic expressions, are not associated with fuzzy sets, but their meanings are related through *possibilistic ontologies* O_i for each domain i , where relations in O_i are modeled by possibility and necessity degrees. For two labels t_i^j and t_i^k :

- $\Pi(t_i^j, t_i^k) = \Pi(t_i^k, t_i^j)$ represents to what extent t_i^j and t_i^k can describe the same thing. A zero possibility means that the two labels never represent the same thing. A positive possibility lesser than 1 expresses that the two terms may mean the same thing, but it is not always the case.
- $N(t_i^j, t_i^k)$ estimates to what extent it is certain that t_i^k is a specialization of t_i^j . Moreover, $N(t_i^j, t_i^k) = N(t_i^k, t_i^j) = 1$ represents genuine synonymy. If t_i^k is a perfect specialization of t_i^j , then $N(t_i^j, t_i^k) = 1$. However, a zero value

only means a total lack of certainty for the specialization relation between these terms. This relation is asymmetric.

These measures must satisfy the following properties:

- Reflexivity: $\Pi(t_i^j, t_i^j) = 1$.
- Symmetry: $\Pi(t_i^j, t_i^k) = \Pi(t_i^k, t_i^j)$.
- $\Pi(t_i^j, t_i^k) \geq N(t_i^j, t_i^k)$, since specialization entails that the meanings overlap.
- $N(t_i^j, t_i^k) > 0 \Rightarrow \Pi(t_i^j, t_i^k) = 1$. If it is somewhat certain that t_i^k specialize t_i^j , then it must be fully possible that they are used for referring to the same thing.
- If labels are precise, we have $N(t_i^j, t_i^j) = 1$. If they are vague, we will suppose $N(t_i^j, t_i^j) \geq \frac{1}{2}$, according to fuzzy pattern matching. This expresses the uncertainty that the query and the data represent “really” the same thing. As an example, two people do not have necessarily the same idea of the price of something found *expensive* by both of them.

The degrees specified in the ontology are actually only defined on a subset of the Cartesian product of the vocabulary $\mathcal{T}_i \times \mathcal{T}_i$. They can be completed using previous properties and the two following forms of transitivity:

$$N(t_i^j, t_i^h) \geq \min \left(N(t_i^j, t_i^k), N(t_i^k, t_i^h) \right), \quad (3)$$

$$\Pi(t_i^j, t_i^h) \geq N(t_i^j, t_i^k) * \Pi(t_i^k, t_i^h). \quad (4)$$

where $*$ is defined as:

$$a * b = \begin{cases} b & \text{if } b > 1 - a, \\ 0 & \text{otherwise.} \end{cases}$$

Equation (3) represents the specialization transitivity [13]. The “hybrid transitivity” (4) states that if t_i^k specializes t_i^j and if t_i^k and t_i^h may refer to the same thing, then the meanings of t_i^j and t_i^h should overlap as well; see [14] for a proof of (3-4).

The certainty degree of the synonymy of t_i^j and t_i^h $Syn(t_i^j, t_i^h)$ can be computed as $\min(N(t_i^j, t_i^h), N(t_i^h, t_i^j))$. Using (3), it can be shown that this degree is max-min transitive. Let $Syn(t, t') = \min(N(t, t'), N(t', t))$. We have $\forall t'', Syn(t, t') \geq \min(Syn(t, t''), Syn(t'', t'))$. Indeed:

$$\begin{aligned} \min(Syn(t, t''), Syn(t'', t')) &= \min(\min(N(t, t''), N(t'', t)), \min(N(t'', t'), N(t', t''))) \\ &= \min(\min(N(t, t''), N(t'', t')), \min(N(t', t''), N(t'', t))) \\ &\leq \min(N(t, t'), N(t', t)) \\ &\leq Syn(t, t') \end{aligned}$$

Therefore, values that are not specified can be deduced from existing ones using the previous properties and relations. Values that cannot be inferred are supposed to be zero. The fact that default possibility is zero corresponds to

a closed world hypothesis, since it is supposed that two terms cannot overlap if it is not specified. From a practical point of view and to simplify the use of the degrees, evaluations will be estimated “at worst”, and the \geq will be generally taken as an equality. It is therefore possible to estimate the relevance degrees between the data and the query even if the searched terms are not directly present in the information representation without using any explicit query expansion stage, as usually proposed in IR.

The extreme binary cases illustrate four situations:

1. the terms are genuine synonyms: $\Pi(t_i^j, t_i^k) = N(t_i^j, t_i^k) = N(t_i^k, t_i^j) = 1$;
2. one of the terms specializes the other term: $\Pi(t_i^j, t_i^k) = N(t_i^j, t_i^k) = 1$ or $\Pi(t_i^k, t_i^j) = N(t_i^k, t_i^j) = 1$;
3. the two meanings overlap, but are not true synonyms nor specializations: $\Pi(t_i^j, t_i^k) = 1$ and $N(t_i^j, t_i^k) = N(t_i^k, t_i^j) = 0$;
4. The meanings are clearly distinct: $\Pi(t_i^j, t_i^k) = N(t_i^j, t_i^k) = N(t_i^k, t_i^j) = 0$.

Intermediary values refine these distinctions. However, only few values in $[0, 1]$ will be generally used, to distinguish the case where a term always certainly specializes another, i.e. $N(t_i^j, t_i^k) = 1$, from the case where it is only *generally* true, which is expressed by $1 > N(t_i^j, t_i^k) > 0$.

As a matter of illustration, Fig. 1 presents a fragment of a simple ontology for accommodation places. This graph is a simplified representation of how an agent may perceive similarity relations between terms. Since it is possible to deduce implicit values for relations between terms, using properties and constraints of the used degrees, only direct links need to be given and are represented here.

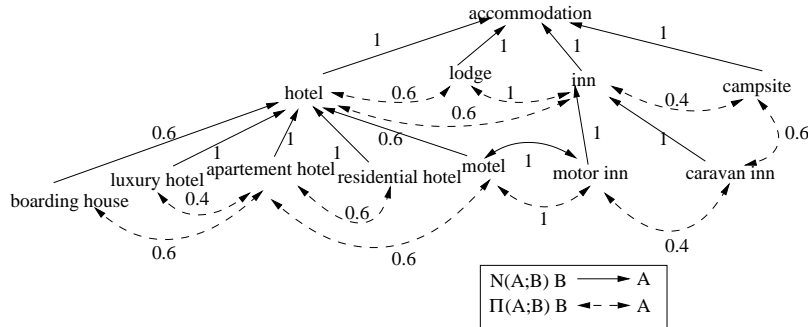


Fig. 1. Accommodation ontology

In Fig. 1, note that words like *lodge* and *inn* are only considered as *possible* synonyms, or as entity that can provide the same services. Nothing can be inferred for the necessity from the possibility degree, and it can exist *lodges* that are not *inns*. On the other hand, both necessity degrees between *motel* and *motor inn* being 1, these terms are considered as genuine synonyms.

The values of degrees present in such ontologies are qualitative in nature, and estimates semantic relations between terms. As an example, $N(\text{hotel}, \text{motel}) = 0.6$ means that it may exist *motels* that are not considered as *hotel*, but that *generally*, *motels* are a kind of *hotels*. Thus, despite the use of numerical values, only the relative ordering between these values is significant, as the purpose is to rank-order query results and not to assess an absolute similarity degree. Practically, only few levels should be used, e.g. $\{0, 0.4, 0.6, 1\}$. These values may be associated for convenience with linguistic labels, such as “very similar”, “rather similar”, etc. to specify the strength of the relations.

Building such ontologies is a complex task, specially if done manually even on a task domain. Ontologies such as WordNet [15] can be used as a starting point. For instance, typed relations used in these ontologies, such as hypernymy can be matched with necessity degrees. Other relations such as “being a part of”, e.g. a *room* and a *hotel*, can be interpreted in terms of possibility degrees. Besides, statistical ontologies can also be built from corpora analysis, extracting relations from terms co-occurrences (e.g. [16]). This may provide a basis for assessing values in possibilistic ontologies using both crisp semantic ontologies and possibilistic rescaling of probabilities (as used in Sect. 5). Subparts of general ontologies can be identified in order to obtain domain specific ontologies.

2.3 Qualitative Pattern Matching

We now consider the case of composed queries. Let Ω be the set of ontologies used to describe the different data domains:

$$\begin{aligned} \Omega &= \{O_i | i = 1, \dots, n\}, \\ O_i &= \{t_i^j \in \mathcal{T}_i\}, \forall i \in \llbracket 1, n \rrbracket, \end{aligned}$$

with \mathcal{T}_i the vocabulary associated with the i^{th} domain. Formally, a piece of information (i.e. a document) is modeled as a set of terms (i.e. keywords), each term in the set belonging to a different ontology:

$$D = \bigwedge_i D_i \text{ and } \exists t'_i \in \mathcal{T}_i, D_i = \{t'_i\}.$$

Queries are conjunctions of disjunctions of (possibly weighted) terms. Thus a query R may be viewed as a conjunction of fuzzy sets R_i representing a conjunction of flexible user needs.

$$R = \bigwedge_i R_i, \text{ where } R_i = \bigvee_j (\lambda_i^j, t_i^j), t_i^j \in \mathcal{T}_i.$$

Note that weighted disjunctions allows us to define new concepts. As an example, a user can specify its own definition of a *cosy lodging* as:

$$(0.5, \text{lodge}) \vee (0.7, \text{motel}) \vee (0.8, \text{apartment hotel}) \vee (1, \text{luxury hotel}).$$

The weights $\lambda_i^j \in [0, 1]$ reflects how satisfactory this term is for the user (i.e. how well it corresponds to his/her request). It is assumed that $\max_j \lambda_i^j = 1$, i.e. at least one query term reflects the exact user requirement. Moreover, importance levels could be introduced between query elements, as described in [7].

The possibilistic query evaluation consists in retrieving all documents D such that the possibility of relevance $\Pi(R, D)$ or the necessity of relevance $N(R, D)$ are non zero. These two relevance degrees are computed as:

$$\Pi(R, D) = \min_i \max_j \min(\lambda_i^j, \Pi(t_i^j, t'_i)) , \quad (5)$$

$$N(R, D) = \min_i \max_j \min(\lambda_i^j, N(t_i^j, t'_i)) . \quad (6)$$

The max parts are weighted disjunctions corresponding to those in the query (where a fuzzy set of more or less satisfactory labels expresses a disjunctive requirement inside the same domain).

In the same way, as the query is a conjunction of elementary requirements pertaining to different domains, the min operator is used in the final aggregation. Note that if R contains a disjunction of redundant terms, that is $R = t \vee t'$ and $N(t, t') = 1$ in the ontology, it can be checked that evaluating t and $t \vee t'$ leads to the same result.

$\Pi(R, D)$ and $N(R, D)$ values estimate to what extent the document D corresponds possibly and certainly to the query R . Results are sorted first using decreasing values of $N(R, D)$, then decreasing values of $\Pi(R, D)$ for pieces of information having the same necessity value.

This matching process can be applied to databases, classical or fuzzy ones, or adapted to information retrieval for collections of sentences or keywords (see Sect. 4), or more generally to documents, as discussed in Sect. 5.

2.4 Other Approaches Using Ontologies

Statistical analysis of texts are used to estimate similarities between terms to define thesauri, maybe interpreted in a fuzzy way, as presented in [17], where fuzzy sets of terms similar to a given term are viewed as representing concepts. These thesauri are used to reformulate queries in order to retrieve more relevant documents. For instance, the *Ontoseek* information retrieval system [18] uses WordNet to expand queries. Ontologies can also be used to index collections [19], as presented in Sec. 5.

In databases, ontologies are used to extend queries to linguistically valued attributes, or to compute similarities between query terms and attribute values. For instance, [20] uses a thesaurus to match queries with a fuzzy database. Imprecise terms are defined as a fuzzy set of terms, and the fuzzy pattern matching is used in the matching process. However, even though links between terms are weighted, this ontology is not a possibilistic one as defined here, and links are sorted as in traditional ontologies.

A recent approach considers relevance rather than similarity between terms [21], where degrees representing terms specialization and generalization are introduced. These degrees are asymmetric, generalization being less favored (from a relevance point of view) than specialization. As an example, *poodle* specializes *dog* at 0.9 whereas *dog* generalizes *poodle* at 0.4. In the approach presented here, two kinds of degrees are used as well, but with a different meaning. The possibility degree is symmetrical, and a positive necessity for $N(t_i^j, t_i^k)$ implies nothing for $N(t_i^k, t_i^j)$, contrary to specialization and generalization degrees of two reversed pairs which are simultaneously strictly positive as in the above example. Moreover, the product used as transitivity operator in [21] leads to a weakening of association weights between terms with the distance in the ontology. Here, the *min* operator implies that the matching is independent of the ontology granularity (inserting a new term between two terms in the ontology cannot change their similarity).

The approach presented in [22] uses ontologies to represent the documents contents, and queries are stated as weighted sets of ontology nodes. Conjunctive queries evaluation is done by comparing minimum subgraphs containing query and document nodes. This comparison is based on a multi-valued degree of inclusion of the document graph in the query graph. Moreover, the documents description takes into account semantic equivalence between expressions, assuming that if a document strongly includes terms, it deals with more general concepts as well, which is equivalent to expand the query with more general terms.

Another approach [23] uses a weighted multilingual ontology to exploit multilingual documents in a translation and search process. In this system, the multilingual ontology is used to translate and expand the query. The query is stated as a subgraph of the ontology, by selecting concepts judged to be relevant. The relevance of each concept can be weighted by the user. The matching is done by computing the inclusion of the query representation in the document representation. A similar approach is presented in [24], where authors use an automatically built fuzzy ontology to expand the user query. The ontology is built using WordNet to extract keywords from a documents collection, the fuzzy relations being computed as in [17]. The ontology is then pruned to eliminate redundant relations.

In [25], a fuzzy ontology is used to summarize news. This ontology is built by fuzzifying an existing ontology using a fuzzy inference mechanism, based on several similarity measures between terms. These measures are computed by textual analysis of corpora. The fuzzy inference inputs are a *part-of-speech distance*, a *term word similarity* that counts the number of common Chinese ideograms in expressions or phrases, and a *semantic distance similarity* based on the distances in the crisp ontology.

Ontologies are also used in [26] to improve clustering of users profiles. These ontologies represent knowledge on the users' domains of interest. The users profiles are then linked through the ontology, which is used to compute a similarity measure between them, allowing a more accurate clustering system.

3 Using Qualitative Pattern Matching on a Database

The qualitative pattern matching framework as previously presented can be applied to classical databases. Possibilistic ontologies, defined on the domain of each linguistic attribute of the database, are thus used to exploit the ontological knowledge about the vocabulary and to make queries on these attributes more flexible. Moreover, as the approach is compatible with the fuzzy pattern matching framework, it can be combined with the evaluation of fuzzy criteria on numerically valued attributes. In this section, some experiment results are presented for illustration purposes. They are carried out on a small but realistic database implemented in the PRETI¹ platform.

Database attribute values are considered as fuzzy ones, and can therefore be represented by linguistic terms associated to fuzzy sets when such a representation exists (e.g. on numerical ranges), or by natural language terms from a known vocabulary stated in a fuzzy ontology as presented in Sec. 2.2. Attribute values represented by terms can be fuzzily matched, exploiting the knowledge from the ontology. Queries will be conjunctions attribute requirements, the conditions for each attribute being stated as a weighted disjunction of acceptable elements of the attribute domain, as presented in Sec. 2.

Description of the PRETI Platform

PRETI¹ is an experimental system used in IRIT laboratory. It contains about 600 records about houses to let in the Aude French department. In this example, only a small subset of the available attributes is used for the sake of simplicity: an identifier, the house location described by one linguistic term, the comfort level encoded by an integer in $\llbracket 0, 4 \rrbracket$, and the price being a real interval giving the minimum and maximum prices.

Used Ontologies

To illustrate the previous framework, fuzzy data and ontologies are needed. Different geographical partitions are used to build the ontology used to represent the location attribute. Districts (French *communes*) are the leafs of the area hierarchies, and are the values actually specified in the database. The sub-ontologies that are used are the following:

cantons: There are 35 *cantons* represented as sets of *communes* (city districts). Their labels start with *c.*. As this classification is crisp, relations in the ontology are pure inclusions ($N = 1$). However, a *canton* and its main *commune* have a reverse necessity of 0.6, since a user may mean a *canton* using the *commune* label. For instance, we have $N(c.limoux, limoux) = 1$ and $N(limoux, c.limoux) = 0.6$, even if the *canton* strictly contains the city.

¹ <http://www.irit.fr/PRETI/accueil.en.php>

arrondissement: the department is split in three administrative districts (French *arrondissement*), described as sets of *cantons*: Carcassonne (*a_carcassonne*), Narbonne (*a_narbonne*) and Limoux (*a_limoux*). These relations are also crisp inclusions ($N = 1$). As for *canton* and *commune*, a reverse necessity is defined between an *arrondissement* and its main *canton*. However, since a user is less likely to use a *canton* as an *arrondissement*, the necessity degree is 0.5 (e.g $N(a_limoux, c_limoux) = 1$ and $N(c_limoux, a_limoux) = 0.5$).

micro-regions: They can be more or less associated to historical or cultural areas, and are also stated as sets of *communes*. Since some *communes* are classified in several micro-regions, the intersection of some micro-regions is not empty. For instance, *narbonne* commune pertains to micro-regions: *cabardes*, *corbieres*, *lauragais*, *narbonnais* and *razes-limouxin*.

Altogether these different administrative districts belong to the same ontology. It is completed by terms referring to physical geography and induced by the fuzzy distinction between mountain, seaside and other areas. The membership of *communes* to these fuzzy terms are given respectively by the average distance to the coast and the average altitude. Moreover, the *pyrenees* micro-region is stated as being included with $N = 0.8$ in *mountain*.

Term mountain

The membership of communes to the *mountain* concept is computed according to their average altitude (between 0 and 1200 meters). The membership function is thus like the one presented in Fig. 2(a), where the “mountainness” of an area is estimated as a percentage of the maximal altitude. It induces the spacial repartition presented in Fig. 2(b). The altitude of communes in black in not known. This will be interpreted as $N = 0, \Pi = 1$, which corresponds to a total lack of information.

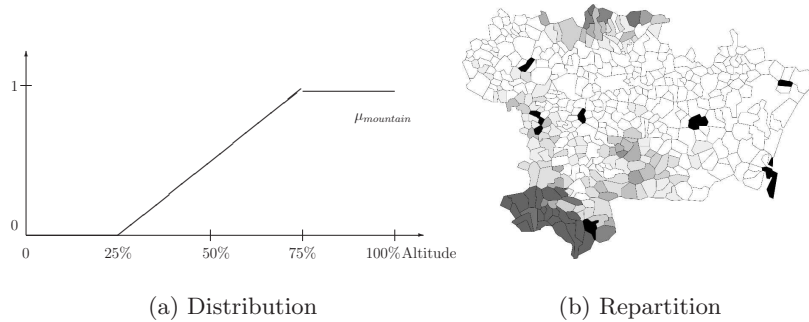
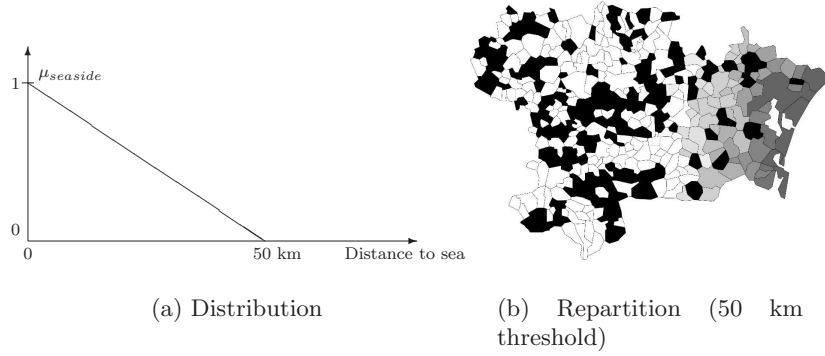


Fig. 2. Term *mountain*

Term seaside

The *seaside* term is defined using the maximum and minimum distance to the coast. Here, if no distance information is given, the degrees are $N = 0$ and $II = 1$ also (the proximity to other communes whose distance to the sea is known is not used). Otherwise, degrees are computed using the membership function described in Fig. 3(a). In the evaluation, a threshold of 50 km has

**Fig. 3.** Term *seaside*

been used; see in Fig. 3(b).

Examples of queries

Let us consider an elementary query: $R_1 = \text{corbieres}$. Evaluating this query on the database without the ontology returns an empty result. Actually, *corbieres* is a micro-region, and the attribute *location* only refers to communes. Using the ontology including micro-regions gives the results reported in Table 1.

Table 1. Results for the query on “*corbieres*”

Number of results	II	N
309	1	1
72	1	0.6
16	1	0

Results with $N = 1$ are due to houses located in *communes* that are included in the searched micro-region. The results at $N = 0.6$ are obtained with a transitivity through *cantons*. For instance, some houses are in the *albies* commune, located in the *cabardes* micro-region. According to the ontology, these terms are not directly related. This degree is obtained in several steps:

1. The micro-regions sub-ontology gives: $N(\text{corbieres}, \text{mouthoumet}) = 1$, since *mouthoumet* is in *corbieres*.
2. The canton sub-ontology gives: $N(\text{mouthoumet}, \text{c_mouthoumet}) = 0.6$ and $N(\text{c_mouthoumet}, \text{albieres}) = 1$, since this commune pertains to this canton.
3. Using transitivity, we thus have $N(\text{corbieres}, \text{albieres}) = 0.6$.

Results having $N = 0$ and $\Pi = 1$ pertain to communes in a micro-region that intersect the *corbieres* micro-region. The global results make sense since the system returns first houses in the requested area, then houses in *cantons* related with the requested micro-region, and lastly houses in a micro-region connected with the requested one (since they intersect it). The size of the area that is put into relation with *corbieres* increases when the relevance decreases, and so does the chance that the retrieved house is in an interesting area for the user. Thus, ontologies allow us therefore to extend the searchable domain of textual attributes without a direct expansion of the query by restating it.

Let us now consider a query that involves preferences, namely $R_2 = \text{pyrenees} \wedge \text{comfort}\{(0.7, 2) \vee 3\}$, expressing that the user is looking for a house to let in Pyrénées, a micro-region, with a comfort level of 3 or possibly 2. The evaluation of the comfort requirement is made in a standard way, namely: $N((0.7, 2) \vee (1, 3), 2) = \Pi((0.7, 2) \vee (1, 3), 2) = 0.7$; $N((0.7, 2) \vee (1, 3), 3) = \Pi((0.7, 2) \vee (1, 3), 3) = 1$. If the ontology is not used, no results are retrieved, whereas using the ontology gives 73 houses (Table 2). The first 13 results pertain to the *pyrenees* micro-region and have a comfort

Table 2. Results for R_2

Number of results	Π	N
13	1	1
47	0.7	0.7
2	1	0.6
7	0.7	0.6
3	1	0
1	0.7	0

of 3. The 47 following houses are also in *pyrenees*, but have a comfort of 2 only. Results with $N = 0.6$ are obtained though their *canton*, as in R_1 , and have a comfort of 3 and 2 respectively, depending on the possibility degree. The following 3 results have a comfort of 3, but are in communes having only a non zero possibility degree of matching with *pyrenees*, since they pertain to cantons having a non empty intersection with *pyrenees*. The last results are in the same situation, but with a comfort level of 2.

Let us now consider the query $R_3 = \text{mountain} \wedge \text{comfort}\{(0.7, 2) \vee 3\}$. The criterion is the same for the comfort attribute, but now uses a fuzzy term for specifying the location. More results are therefore retrieved (see Table 3).

Table 3. Results for R_3

Number of results	Π	N
7	1	1
1	1	0.9
7	1	0.8
55	0.7	0.7
2	1	0.6
2	0.7	0.6
2	1	0.5
1	0.7	0.5
130	1	0.4
151	0.7	0.4
2	0.7	0.3
8	1	0.2
2	0.7	0.2
3	1	0.1
9	0.7	0.1
5	1	0
8	0.7	0

In the ontology, $N(\textit{mountain}, \textit{pyrenees}) = 0.8$. The first group of 13 houses found with R_2 are now discriminated into the three first groups of results of Table 3. The two first groups pertains to *mountain* with a necessity $N \geq 0.9$ and with a comfort of 3. The 7 houses with $N = 0.8$ either pertains to *pyrenees* or have a necessity with *mountain* at least equal to 0.8 given by the altitude. The same analysis can be done for other results. The possibility value 0.7 is induced by a comfort of 2, as previously. For necessity degrees with a value less than 0.7, the final degree value is determined by the inclusion degree of the commune in *mountain*, since it is the minimum of this value and the one implied by the comfort level, which is at least 0.7. The system returns here 395 houses.

In this case, the use of a fuzzy term, with a broader sense than the area name (*pyrenees*) leads to more results, and provides a better granularity in the result ordering.

Let us consider another query involving disjunction between symbolic labels, namely $R_4 = (\textit{seaside} \vee \textit{c.carcassonne}) \wedge [0, 1500]$ (a house near the coast or in the carcassonne *canton* and with a price lower than 1500).

The query price requirement is evaluated in the following way: $\Pi = N = 1$ if the attribute value interval is included in the query one; $\Pi = 1, N = 0$ if the two intervals overlap; $\Pi = 0, N = 0$ if the two intervals are disjoint.

R_4 leads to the results presented in Table 4. Most of the non-integer values for the necessity degrees are induced by the evaluation of the term *seaside*, since the requirement $\textit{c.carcassonne} \wedge [0, 1500]$ leads only to results with $\Pi = 1, N = 0$ and $\Pi = 1, N = 0.5$ due to a transitivity through *arrondissement*. Again, the presence of a fuzzy term in the ontology leads to a more refined

Table 4. Results for R_4

Number of results		
II	N	
1	1	1
2	1	0.9
3	1	0.8
1	1	0.7
7	1	0.6
36	1	0.5
1	1	0.4
8	1	0.3
237	1	0

ranking. This is even more effective when using preference weights in the disjunction of symbolic terms as in the following query.

By weighting *carcassonne*, $R_5 = (seaside \vee (0.7, c_carcassonne)) \wedge [0, 1500]$ states that the user prefers an house closed to the coast, even if one in the historical city of Carcassonne would be still acceptable (Carcassonne is far from the coast, $N(seaside, c_carcassonne) = 0$). The difference between R_4

Table 5. Results for R_5

Number of results		
II	N	
1	1	1
2	1	0.9
3	1	0.8
1	1	0.7
7	1	0.6
17	1	0.5
19	0.7	0.5
1	1	0.4
8	1	0.3
184	1	0
53	0.7	0

and R_5 is introduced by the weighting of *carcassonne*. In Table 5, the 36 houses that had $II = 1, N = 0.5$ in Table 4 are discriminated in 17 results with $II = 1, N = 0.5$, which correspond to houses with a 0.5 degree with *littoral*, and 19 with $II = 0.7, N = 0.5$, corresponding to houses in the Carcassonne *canton*. In the same way, the 237 last houses, of which 53 are obtained through a relation with *c_carcassonne*, have now a possibility reduced to 0.7 due to the weighting in the query.

Qualitative pattern matching with databases containing linguistic labels, allows the semantic evaluation of flexible queries that also use linguistic terms. The evaluation of the semantic similarity of terms is done by means of possibilistic ontologies, but may also use fuzzy set based representations, especially

for terms referring to numerical scales, since the approach is fully compatible with standard fuzzy pattern matching. As shown by the above illustration, the evaluation process does not look for a strict matching between identical terms, which avoids the reformulation of the query.

These ideas apply to information retrieval as well, since the data are then textual. This is the topic of the next sections.

4 Retrieving Titles Using Qualitative Pattern Matching

In this illustration, a collection of titles of articles is considered. Titles are viewed as set of keywords, obtained by lemmatizing their significant terms and forgetting the stop-words. Therefore, the information does no longer refer to distinct attributes (with their own domain) as in the database example of the previous section. Keywords correspond to a unique multiple-valued attribute, in which terms pertain to the same global domain \mathcal{T} .

In the following, it is assumed that all the terms used in the query and in the titles are in the ontology. This can be practically achieved by enforcing the user to choose query terms in the ontology, and by making sure that terms appearing in representative titles are indeed in the ontology.

Queries are still conjunctions of disjunctions of possibly weighted terms, but all terms are now in the same domain (i.e. vocabulary). Moreover, weights are also introduced at the conjunctions level in order to express the relative importance of the elementary requirements of the query. Queries are thus weighted Boolean expressions on keywords. Namely,

$$D = \{t'_i, t'_i \in \mathcal{T}\}, R = \bigwedge_k \left(\omega_k, \bigvee_j (\lambda_k^j, t_k^j) \right), \text{ with } t_k^j \in \mathcal{T}.$$

This *importance* weighting obey the same constraint as the weights λ_k^j . In practice, disjunctions are between terms which are more or less interchangeable for the user (the weight λ_k^j expressing his/her preference between them). The weight ω_k expresses how compulsory is each elementary requirement in the conjunction.

In this fuzzy context, conjunctions are still evaluated by the *min* operator, and disjunctions by the *max* operator to compute the possibility and necessity degrees. The evaluation equations (5)-(6) are therefore rewritten as:

$$H(R, D) = \min_k \max_{i,j} \min(\lambda_k^j, H(t_k^j, t'_i)), \quad (7)$$

$$N(R, D) = \min_k \max_{i,j} \min(\lambda_k^j, N(t_k^j, t'_i)), \quad (8)$$

to acknowledge the multiple-valued aspect of the keywords attribute. Since the ω_k 's are importance weights, this leads to the more general weighted min formula (the above formulas (7-8) are retrieved when all ω_k are equal to 1):

$$\begin{aligned} \Pi(R, D) &= \min_k \max \left(1 - \omega_k, \max_{i,j} \min(\lambda_k^j, \Pi(t_k^j, t_i')) \right), \\ N(R, D) &= \min_k \max \left(1 - \omega_k, \max_{i,j} \min(\lambda_k^j, N(t_k^j, t_i')) \right). \end{aligned}$$

Therefore, having an importance weight ω_k less than 1 leads to retrieve results violating the corresponding elementary requirement with a degree at most equal to $1 - \omega_k$.

Experimentation protocol

In this experimentation, the TREC protocol is followed, defining queries and their corresponding relevant documents to compute precision values for the retrieval system. However, the experiments reported below are not a real evaluation, as in TREC campaigns, but rather an illustration of the potentials of the approach and its application to textual information retrieval.

4.1 Data Description

The collection contains about 200 titles of computer science articles, mainly in English but some in French, from artificial intelligence and information retrieval fields. In order to index these “documents” and to evaluate queries using qualitative pattern matching, a simple ad hoc ontology corresponding to titles terms is used. The ontology has been built a posteriori to fulfill the assumption that all documents and queries terms must be in the ontology. First, terms are generalized by their stem, given by the Porter algorithm ($N(stem, term) = 1$). To represent the few cases where different terms lead to the same stem, this relation is not considered as genuine synonymy, and the reverse necessity is set to $N(term, stem) = 0.9$. Other relations are introduced by translating French terms in English and considering a term and its translation as synonyms. Moreover, some compound expressions, such as *fuzzy set*, as well as the weights associated to the links between the terms, are added manually. A fragment of this ontology is shown in Fig. 4.

4.2 Examples of Queries

Queries used in the illustration are:

1. $nutrition \vee (repas \wedge \acute{e}quilibr\acute{e})$, reformulated as $(nutrition \vee repas) \wedge (nutrition \vee \acute{e}quilibr\acute{e})$ to fit the query format.
2. $(nutrition \vee meal) \wedge (nutrition \vee balanced)$, which is a translation of the previous one.
3. $fuzzy \wedge information$
4. $model \wedge (reasoning \vee decision)$

To help interpreting the precision values presented in the following, the number of relevant documents to each query is given in Table 6.

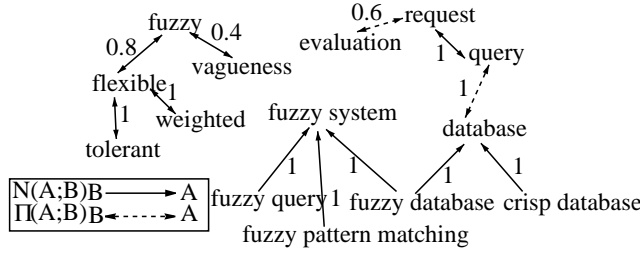


Fig. 4. Ontology fragment for the titles collection

Table 6. Number of relevant documents for each query

Query	Relevant doc.
1	6
2	6
3	15
4	41

4.3 Evaluation and Results

In this experiment, results obtained by just looking if each query term is present or not in the titles, are compared with those obtained by using the ontology. Tables 8(a) and 8(b) show the precision values of the two evaluations respectively.

Table 7. Precision values for queries the evaluations

Query	P5	P10	P15	AvgPr	Query	P5	P10	P15	AvgPr
1	1.00	0.50	0.33	0.83	1	1.00	0.60	0.40	1.00
2	1.00	0.50	0.33	0.83	2	1.00	0.60	0.40	1.00
3	0.20	0.10	0.07	0.07	3	0.40	0.60	0.53	0.34
4	1.00	0.70	0.47	0.17	4	1.00	0.90	0.87	0.39

(a) without ontology

(b) using ontology

For the two first queries, one of the relevant documents is in French. In the first case, the system cannot translate the query terms to match the index and does not retrieve the French document that does not contains *nutrition* (title 218, see Table 8). Only 5 among the 6 relevant documents are retrieved. that is 5 over 6, thus the P5 value of 1 and P10 of 0.5. As *nutrition* has the same writing in English as in French, the 5 other documents are retrieved. On the other hand, the translation is possible using the ontology, and the

system retrieves all relevant documents, thus both average and P5 precisions are equal to 1. Relevant titles for the two first queries are given in Table 8.

Table 8. Relevant titles for queries 1 and 2

135	Nutri-Expert, an Educational Software in Nutrition
218	Nutri-Expert et Nutri-Advice, deux logiciels d'aide à la construction de repas équilibrés pour l'éducation nutritionnelle
234	Balancing Meals Using Fuzzy Arithmetics and Heuristic Search Algorithms
237	Multicenter randomized evaluation of a nutritional education software in obese patients
238	Expert system DIABETO and nutrition in diabetes
239	Evaluation of microcomputer nutritional teaching games in 1876 children at school

Without ontology, query 3 selects only one title: “Fuzzy sets and fuzzy information granulation theory”, whereas with the ontology this query retrieves 9 more titles presented in Table 9. In the ontology, *possibilistic logic* IS A kind

Table 9. Detailed results for query 3 *with* ontology

Doc. #	I	N	Title
223	1	1	Fuzzy sets and fuzzy information granulation theory
259	1	1	Quasi-possibilistic logic and its measures of information and conflict
264	1	1	Practical Handling of Exception-tainted rules and independence information in possibilistic logic
287	1	0.8	Fuzzy logic techniques in multimedia database querying
156	1	0.8	Fuzzy logic techniques in Multimedia database querying: a preliminary investigation of the potentials
129	1	0.8	Flexible queries in relational databases - The example of the division operator
114	1	0.8	Semantics of quotient operators in fuzzy relational databases
236	1	0.5	Fuzzy scheduling: Modelling flexible constraints vs. coping with incomplete knowledge
216	1	0.5	Uncertainty and Vagueness in Knowledge-Based Systems
137	1	0.5	Checking the coherence and redundancy of fuzzy knowledge bases
195	0.5	0	Handling locally stratified inconsistent knowledge bases
128	0.5	0	Some syntactic approaches to the handling of inconsistent knowledge bases

of *fuzzy logic*, which IS *fuzzy*, and therefore, $N(\text{fuzzy}, \text{possibilistic logic}) = 1$ (title 264). Moreover, the ontology considers *flexible* and *fuzzy* as 0.8 synonyms, as well as *data* and *information*; *database* being a specialization of *data*, giving degrees for titles from 287 to 114 in Table 9. The 0.5 degree is given by a 0.5 necessity between terms *information* and *knowledge* in the ontology. Note that titles that are not closely related to the query, but still

weakly relevant, such as titles 195 and 128 are also retrieved at the end of the list, but with a positive possibility degree only (the possibility weights come from the application of (4)). However, since the qualitative pattern matching does not require a perfect match between the query and the data, more relevant titles are retrieved and the precision for this query is thus improved.

Weighted Queries

Consider now weighted versions of the previous queries 3 and 4:

3. $(0.3, fuzzy) \wedge (1, information)$
4. $(0.7, model) \wedge (1, ((0.8, reasoning) \vee (1, decision)))$

For query 3, greater importance is given to the term *information* w.r.t. *fuzzy*, while in query 4, weight 0.7 expresses that retrieving *model* or an equivalent term in the title is less important than satisfying the disjunctive part of the query. Weight 0.8 reflects lower preference for *reasoning* compared to *decision*. Results of these queries are presented in Tables 10 and 12 respectively.

Table 10. Results of weighted queries *without* ontology

Query	P5	P10	P15	AvgPr
3	0.20	0.30	0.20	0.11
4	1.00	1.00	0.87	0.73

Without using the ontology, results of the weighted version of query 3 are improved as the average precision raises from 0.07 to 0.11. The non-weighted version retrieved only one document. Lowering the importance of *fuzzy* leads to retrieve more titles as detailed in Table 11

Even if quite a lot of documents have a debatable relevance (average precision 0.11), a few more relevant ones are retrieved. This improvement is due to the collection itself that contains mainly documents on *fuzzy* topics. Therefore, this term is not as discriminant as *information* in this particular collection, and its importance in the query can be lowered.

In the same way, for R_4 , more relevant documents are retrieved in the 10 first results due to the lower importance of *model*, thus the increasing of P10.

The weighting of query terms allows therefore to exploit knowledge about the collection, lowering the importance of terms known not to be discriminant.

The impact of the combined use of the weights and the ontology is difficult to analyze on this small experiment. Query 4 performance is improved by both the weighting and the ontology. Nevertheless, weighting *fuzzy* in query 3 with the ontology decreases the precision. Indeed, a 0.3 weight leads to retrieve titles that are not linked with the concept of fuzziness, but having possibility and necessity degrees of 0.7 (which correspond to the fact that it is not very important to have *fuzzy* in the title). Therefore, they obtain a better rank than titles linked with this concept with a 0.5 necessity degree, which would

Table 11. Results of weighted query 3 *without* ontology

223	1	1	Fuzzy sets and fuzzy information granulation theory
292	0.7	0.7	Internet-based information discovery: Application to monitoring science and technology
291	0.7	0.7	TétraFusion: Information Discovery on the Internet
288	0.7	0.7	Information discovery from semi-structured sources Application to astronomical literature
284	0.7	0.7	On using genetic algorithms for multimodal relevance optimisation in information retrieval
264	0.7	0.7	Practical Handling of Exception-tainted rules and independence information in possibilistic logic
260	0.7	0.7	On the use of aggregation operations in information fusion processes
259	0.7	0.7	Quasi-possibilistic logic and its measures of information and conflict
251	0.7	0.7	Logical representation and fusion of prioritized information based on guaranteed possibility measures: Application to the distance-based merging of classical bases
193	0.7	0.7	Possibilistic merging and distance-based fusion of propositional information

Table 12. Results of fuzzy queries using ontology

Query	P5	P10	P15	AvgPr
3	0.40	0.30	0.27	0.34
4	1.00	1.00	1.00	0.76

actually be more relevant. This raises the issue of the commensurateness of the scales used for assessing the weights in the query and in the ontology.

Generally, the implicit query expansion by means of the ontology improves the system performances. This simple illustration shows that the ontology is an important aspect of the system efficiency, and therefore that this approach depends on the quality of the ontology used. By introducing weights in the query, the user can represent its preferences and priorities, as well as take into account some knowledge about the collection. However, the impact of the weighting is a difficult aspect to evaluate, since some values can improve results and some can worsen them, specially when using the ontology which gives a additional fuzzification of the final relevance degree. This experiment cannot be considered as an evaluation of a real system, due to the limited number of titles in the collection and the few queries used. However, it illustrates the approach, showing its possibilities and limitations.

5 Toward an Extension of the Approach to Full-text IR

In the previous section, even if no database attribute is considered, the illustration cannot be seen as genuine full-text IR. Indeed, no statistical analysis is done to compute terms importance in the document. In this section, possibilities of extension of the model to full-text IR, by using statistical analysis to estimate possibility and necessity degrees between the ontology terms and the documents are explored.

5.1 Possibilistic Indexing

To be homogeneous with the ontology model, the association between documents and the ontology nodes must be stated using the same possibility and necessity degrees, taking into account the statistical weights of the terms in the documents. Classically, the significance weight ρ_i^j for a given term t_i w.r.t. a document D_j is computed by combining its frequency tf_{ij} in D_j and its inverse frequency $idf_i = \log(d/df_i)$, where df_i is the number of documents containing t_i and d is the number of documents in the collection. The weights ρ_i^j are assumed to be rescaled between 0 and 1. The document D_j is therefore represented by the fuzzy set of its significant terms [4, 27]: $D_j = \{(\rho_i^j, t_i), i = 1, n\}$, where n is the number of terms in the ontology.

Assuming that the ρ_i^j is an intermediary degree between the possibility and the necessity that the term describes the document, the possibility and necessity degrees can be computed as follows [28]:

$$\begin{cases} \Pi(t_i, D_j) = 2\rho_i^j ; N(t_i, D_j) = 0 & \text{if } \rho_i^j < \frac{1}{2} , \\ \Pi(t_i, D_j) = 1 ; N(t_i, D_j) = 2\rho_i^j - 1 & \text{otherwise .} \end{cases} \quad (9)$$

The intuition underlying (9) is that a sufficiently frequent term in the document is necessarily somewhat relevant, while a less frequent term is only possibly relevant.

The ontology model agrees with the *synset* concept in WordNet. A synset is a set of synonymous terms such as: $S = \{t_i \in \mathcal{T}\}$ such that $\forall(i, j), t_i, t_j \in S \iff t_i \neq t_j, \Pi(t_i, t_j) = 1$ and $N(t_i, t_j) = N(t_i, t_j) = 1$. This allows us to take synsets as ontology nodes, thanks to the transitivity properties (3)-(4). Indeed, a term (used in a given sense) belongs to only one synset and is a synonym of all other synset terms. We have to estimate to what extent a synset $S = \{t_i, i = 1, p\}$ describes a document D_j , that is to compute $\Pi(S, D_j)$ and $N(S, D_j)$. Since all synset terms are supposed to describe the document equally, we have $\Pi(S, D_j) = \max_{i, t_i \in S}(\Pi(t_i, D_j))$ and $N(S, D_j) = \max_{i, t_i \in S}(N(t_i, D_j))$ (see Fig. 5).

Indexing example:

As an example, let us consider the document D represented by the index given in Table 13.

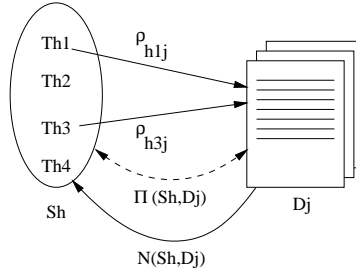


Fig. 5. Links between document and synset

Table 13. Index example

Term	ρ	Π	N
Database	0.6	1	0.2
Artificial Intelligence	0.2	0.4	0
AI	0.7	1	0.4
Machine learning	0.8	1	0.6

This suggests that this document deals with artificial intelligence, more specially with machine learning, applied to databases. Notice that despite *artificial intelligence* and *AI* have exactly the same meaning, their weights are different, since from a statistical point of view, the term *AI* is more frequent in the document than *artificial intelligence*. Thus, the (Π, N) degrees between the synset $\{ArtificialIntelligence, AI\}$ and D is $(\Pi, N) = (\max(0.4, 1), \max(0, 0.4)) = (1, 0.4)$.

5.2 Query Evaluation

Given a collection of documents indexed using an ontology, the query evaluation can be done similarly as described in Sect. 5. However the significance degrees between query terms and documents are no longer supposed to be 1 or 0. Taking into account the possibility and the certainty of significance as given by (9), leads for a query R and a document D , to the following relevance status value (rsv):

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

where degrees are given by:

$$\begin{aligned} \Pi(R, D) &= \min_k \max_{i,j} \min(\lambda_k^j, \Pi(t_k^j, t_i), \Pi(t_i, D)) , \\ N(R, D) &= \min_k \max_{i,j} \min(\lambda_k^j, N(t_k^j, t_i), N(t_i, D)) , \end{aligned}$$

The importance weights of the elementary requirement in a query (ω_k) can be added, which leads to:

$$\begin{aligned}
H(R, D) &= \min_k \max \left(1 - \omega_k, \max_{i,j} \min(\lambda_k^j, H(t_k^j, t_i), H(t_i, D)) \right), \\
N(R, D) &= \min_k \max \left(1 - \omega_k, \max_{i,j} \min(\lambda_k^j, N(t_k^j, t_i), N(t_i, D)) \right).
\end{aligned}$$

The above expressions provide bases for the extension of the approach to general documents information retrieval.

Besides, in the above formulas, the aggregation of the evaluations associated with each elementary requirement is performed by means of the conjunction *min*. It is well known in information retrieval that the minimum operation is often too restrictive in practice, and is usually outperformed by other operations such as the sum. However, it has been shown in a recent work [29], that it is possible to refine the minimum operation (using a leximin ordering on ordered sets of values to be compared), and to obtain results as good or even better than with the sum. Such a refinement could be applied also in the above approach.

6 Conclusion

The approach described in this chapter is an adaptation of fuzzy pattern matching to purely linguistic terms. The main idea is to retrieve information containing terms that may not match exactly those of the query. To cope with this point, a *possibilistic ontology* is used, where the relations between terms are stated by the possibility and the certainty that their meanings refer to the same thing. This allows us to specify semantic relations, such as synonymy or specialization and generalization of meanings. Thanks to the transitivity properties of possibilistic ontologies, relations that are not explicitly stated can be deduced. A property of this model is the independence of the similarity with respect to the hierarchical distance of terms in the ontology, and therefore to the granularity of the vocabulary.

The application of qualitative pattern matching to databases allows the evaluation of flexible queries on linguistic terms, in agreement with the more standard handling of queries and data represented by fuzzy sets. Its use in information retrieval systems, avoiding query reformulation owing to the a priori vocabulary knowledge contained in the ontology. Since the matching is qualitative in nature, results can be rank-ordered even though no document matches exactly the query. The experiments undertaken separately in a database and in a textual data collection show that the approach is viable for both fields. Indeed, results are improved by the use of the possibilistic ontology and prioritized queries.

References

1. G. Grefenstette, editor. *Cross-Language Information Retrieval*. Kluwer Academic, Boston, 1998.
2. G. Bordogna and G. Pasi. A fuzzy linguistic approach generalizing boolean information retrieval: a model and its evaluation. *Journal of the American Society for Information Science*, 44(2):70–82, 1993.
3. T. Andreasen, H. Christiansen, and H. L. Larsen, editors. *Flexible Query Answering Systems*. Kluwer, 1997.
4. D. Kraft, G. Bordogna, and G. Pasi. Fuzzy set techniques in information retrieval. In *Fuzzy Sets in Approximate Reasoning and Information Systems*, chapter 8, pages 469–510. Kluwer Academic Publishers, 1999.
5. D. Dubois, H. Prade, and C. Testemale. Weighted fuzzy pattern matching. *Fuzzy Sets and Systems*, 28:313–331, 1988.
6. M. Boughanem, Y. Loiseau, and H. Prade. Graded pattern matching in a multilingual context. In *Proc. 7th Meeting Euro Working Group on Fuzzy Sets*, pages 121–126. Eurofuse, Varena, 2002.
7. Y. Loiseau, H. Prade, and M. Boughanem. Qualitative pattern matching with linguistic terms. *Ai Communications, The European Journal on Artificial Intelligence (AiCom)*, 17(1):25–34, 2004.
8. G. Salton. Experiments in automatic thesaurus construction for information retrieval. In *IFIP Congress*, pages 115–123, 1971.
9. M. Cayrol, H. Farreny, and H. Prade. Fuzzy pattern matching. *Kybern.*, 11:103–16, 1982.
10. D. Dubois and H. Prade. Tolerant fuzzy pattern matching: an introduction. In P. Bosc and J. Kacprzyk, editors, *Fuzziness in Database Management Systems*, pages 42–58. Physica-Verlag, 1995.
11. P. Resnik. Semantic similarity in a taxonomy: an information-based measure and its application to problem of ambiguity in natural language. *J. Artif. Intellig. Res.*, 11:95–130, 1999.
12. A. Bidault, C. Froidevaux, and B. Safar. Similarity between queries in a mediator. In *Proc. 15th European Conference on Artificial Intelligence*, pages 235–239. ECAI’02, Lyon, July 2002.
13. J.P. Rossazza, D. Dubois, and H. Prade. A hierarchical model of fuzzy classes. In R. De Caluwe, editor, *Fuzzy and Uncertain Object-Oriented Databases*, pages 21–62. World Pub. Co., 1997.
14. D. Dubois and H. Prade. Resolution principles in possibilistic logic. *Int. Jour. of Approximate Reasoning*, 4(1):1–21, 1990.
15. G.A. Miller, R. Beckwith, C.Fellbaum, D. Gross, and K.J. Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244, 1990.
16. C.J. Crouch. An approach to the automatic construction of global thesauri. *Information Processing and Management*, 26(5):629–640, 1990.
17. S. Miyamoto. *Fuzzy sets in Information Retrieval and Cluster Analysis*. Kluwer Academic Publisher, 1990.
18. N. Guarino, C. Masolo, and G. Vetere. Ontoseek : content-based access to the web. *IEEE Intelligent Systems*, 14(3):70–80, 1999.
19. Mustapha Baziz, Mohand Boughanem, Nathalie Aussenac-Gilles, and Claude Chriment. Semantic cores for representing documents in ir. In *SAC’2005- 20th*

- ACM Symposium on Applied Computing*. Santa Fe, New Mexico, USA., 13-17 mars 2005.
20. N. Mouaddib and P. Subtil. Management of uncertainty and vagueness in databases: the FIRMS point of view. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 5(4):437–457, 1997.
 21. H. Bulskov, R. Knappe, and T. Andreassen. On measuring similarity for conceptual querying. In *Flexible Query Answering Systems, LNAI 2522*, pages 100–111. Springer, 2002.
 22. M. Boughanem, G. Pasi, and H. Prade. Fuzzy set approach to concept-based information retrieval. In *10th International Conference IPMU*, pages 1775–1782. IPMU'04, Perugia (Italy), July 2004.
 23. V. Cross and C.R. Voss. Fuzzy ontologies for multilingual document exploitation. In *Proc. of the 18th Conference of NAFIPS*, pages 392–397. New York City, IEEE Computer Society Press, June 1999.
 24. D.H. Widyantoro and J. Yen. A fuzzy ontology-based abstract search engine and its user studies. In *FUZZ-IEEE*, pages 1291–1294, 2001.
 25. C-S Lee, Z-W Jian, and L-K Huang. A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man and Cybernetics*, 35(5):859–880, October 2005.
 26. A. Smirnov, M. Pashkin, N. Chilov, T. Levashova, A. Krizhanovsky, and A. Kashaevnik. Ontology-based user and requests clustering in customer service management system. In V. Gorodetsky, J. Liu, and V. Skormin, editors, *Autonomous Intelligent Systems: Agent and Data Mining*, pages 231–246. Int. Workshop , AIS-ADM 2005, Springer-Verlag, 2005.
 27. D.A. Buell. An analysis of some fuzzy subset applications to information retrieval systems. *Fuzzy Sets and Systems*, 7(1):35–42, 1982.
 28. H. Prade and C. Testemale. Application of possibility and necessity measures to documentary information retrieval. *LNCS*, 286:265–275, 1987.
 29. M. Boughanem, Y. Loiseau, and H. Prade. Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. In *3rd International Workshop on Adaptive Multimedia Retrieval*. AMR'05, Glasgow (UK), July 2005.