

# Graded pattern matching in a multilingual context

**Mohand Boughanem**

IRIT, Université Paul Sabatier  
boughane@irit.fr

**Yannick Loiseau**

IRIT, Université Paul Sabatier  
loiseau@irit.fr

**Henri Prade**

IRIT, Université Paul Sabatier  
prade@irit.fr

## Abstract

‘Fuzzy pattern matching’ (FPM) has been in use for two decades, in particular for querying fuzzy databases. FPM returns two matching degrees: for each piece of fuzzy data in the base, the possibility and the certainty that it matches a pattern representing a flexible request, are computed. The same idea has been suggested in information retrieval for distinguishing between keywords which certainly apply for describing a document, and those which are only optional. A counterpart of FPM has been recently proposed for estimating levels of matching between a request and pieces of data expressed with words which are related through a qualitative ontology, where approximate synonymy and specialization relations are encoded in terms of certainty and possibility degrees respectively. The paper outlines an extension and an adaptation of these ideas to the retrieval of multilingual documents.

## 1 Introduction

Querying processes in information systems may take into account user’s preferences, by allowing for flexible queries (e.g. [1]), which then enables the system to rank-order results. Fuzzy set-based approaches have been developed for representing flexible queries, and can be applied to regular as well as fuzzy databases. A tool, called ‘fuzzy pattern matching’ (FPM)[4, 6] has been proposed in

the framework of possibility theory, which computes to what extent it is possible and certain that a piece of information satisfies a flexible request expressed by means of fuzzy sets representing the user’s preferences. This idea can be also extended to information retrieval. Documents are then described by keywords which are more or less certainly, or only possibly, relevant to some extent. Similarly, the request may involve keywords which are more or less compulsory, as well as others which are only optional [12].

In FPM, each label appearing in the request or in the database is represented by a fuzzy set. Fuzzy sets defined on the same attribute domain can be compared, by means of possibility and necessity measures. Recently, a qualitative approach has been proposed where these features are adapted to handle symbolic labels [10]. The intended purpose of the approach is to deal with queries stated in terms of linguistic labels, which may be weighted for expressing preferences. The matching between the labels in the request and the data is a matter of semantic similarity computed by means of a weighted network associated with each attribute domain. Thus, the labels are no longer explicitly associated with fuzzy set representations, but their semantic relationships are still assumed to be estimated in terms of the two above-mentioned measures.

The purpose of this paper is to apply the above ideas to multilingual information retrieval where documents are represented by extracted terms and request involves weighted conjunctions and/or disjunctions of keywords. Section 2 provides a short background on FPM. Section 3 summarizes the handling of symbolic labels in qualitative pattern matching. Section 4 discusses its

application to multilingual information retrieval and illustrates it on an example.

## 2 Fuzzy Pattern Matching

A pattern is a set of elementary requirements encoded by labels of properties referring to attribute domains. The basic idea is to attach to each label of a pattern the membership function of a fuzzy set restricting the values which are more or less compatible with the meaning of the label. These values belong to the domain of the attribute which the label refers to. For instance, the label *tall* refers to a scale of heights, and corresponds to a fuzzy subset of this domain, which depends on the context. In place of a numerical domain, we may have a discrete set of elements as well. Besides, data are also represented by lists of labels whose components are associated with fuzzy sets. These fuzzy sets are viewed as possibility distributions which model the imprecision pervading the data, and restrict the more or less possible values of the considered attributes. Such lists contain possibly ill-known attribute values pertaining to the description of objects. Namely, a component in a list refers to only one (ill-located) value of the domain of the concerned attribute (which is supposed to be single-valued).

Thus, a fuzzy pattern represents an imprecisely described class of objects which are looked for. Namely, let  $T$  and  $T'$  be respectively a pattern label (i.e. a requirement) and an item component pertaining to the same single-valued attribute (i.e. a piece of data), which are to be compared.  $T$  and  $T'$  refer to fuzzy sets of the same domain  $U$  conveying their meanings. Let  $\mu_T$  be the membership function associated to label  $T$  and  $\pi_{T'}$  be the possibility distribution attached to  $T'$ . Both are mappings from  $U$  to  $[0, 1]$ . Let  $u$  be an element of  $U$ . Then  $\mu_T(u)$  is the grade of compatibility between the value  $u$  and the meaning of  $T$ . Namely,  $\mu_T(u) = 1$  means full compatibility with  $T$  and  $\mu_T(u) = 0$  means total incompatibility with  $T$ . By contrast,  $\pi_{T'}(u)$  is the grade of possibility that  $u$  is the value of the attribute describing the considered item.  $T'$  is a fuzzy set of *possible* values (only one of which is the genuine value of the ill-known attribute), while  $T$  is a fuzzy set of values *more or less* compatible with

the preferences expressed by the user. For instance,  $\pi_{T'}(u) = 1$  means that  $u$  is totally possible (there may exist distinct values  $u$  and  $u'$  such as  $\pi_{T'}(u) = \pi_{T'}(u') = 1$ ), while  $\pi_{T'}(u) = 0$  means that  $u$  is totally impossible as an attribute value of the object to which the item pertains. Here,  $\mu_T$  and  $\pi_{T'}$  are supposed to be normalised, i.e. there is a value which is totally compatible with  $T$ , and a value totally possible in the range  $T'$ . Two scalar measures estimate the compatibility between a pattern element  $T$  and its counterpart  $T'$  in the considered piece of data, namely a degree of possibility of matching  $\Pi(T; T')$  and a degree of necessity of matching  $N(T; T')$  respectively defined by [4]:

$$\begin{aligned} \Pi(T; T') &= \sup_{u \in U} \min(\mu_T(u), \pi_{T'}(u)) \text{ and} \\ N(T; T') &= \inf_{u \in U} \max(\mu_T(u), 1 - \pi_{T'}(u)). \end{aligned}$$

The basic asymmetry of the pattern-data matching is preserved by the second measure, since  $N$  computes a degree of inclusion between fuzzy sets, while  $\Pi$  estimates the non-emptiness of an intersection. The atomic measures of possibility and necessity are aggregated separately in order to obtain two global measures between the whole pattern and the whole item. When the pattern expresses a conjunction of elementary requirements " $T_1$  and  $\dots$   $T_n$ ", this aggregation is performed using the min operation, and preserves the respective semantics of the measures in terms of possibility and necessity [6].

## 3 Ontology-Based Matching

FPM estimates the possibility and the necessity that the meanings of two labels represented by fuzzy sets coincide. However, such a representation is not always available for computing matching degrees. Measures of semantic similarity between words have been thoroughly studied in the information retrieval literature, taking advantage of distances between nodes in a taxonomy, or based on common probabilistic information content (e.g. [13]). Alternatively, one commonly investigated strategy when a user's query fails (because data encoded in perfectly identical terms cannot be retrieved), is to generate approximately similar queries in place of the initial one (e.g [2]) on the basis of ontologies. Generally speaking,

these concerns could also be related to the “computing with words” research trend [14]. Keeping inspiration from FPM, we now summarize a recent approach for evaluating approximate similarities, based on semantic networks weighted in terms of possibility and necessity degrees.

Let’s consider a database whose items are described by a set of identified attributes  $i = 1, n$ . Let  $\mathcal{T}_i$  be the vocabulary relative to the attribute  $i$ . Let us assume first that each attribute value  $T'_i$  is given by a single label or term  $t'_{ij}$ , i.e.  $T'_i = \{t'_{ij}\}$ . Since each attribute contributes to the information description, a piece of data is a conjunction of labels, which can be symbolically written:  $T' = \bigwedge_{i \in [1;n]} T'_i$ .

In the same way, requests are conjunctions of weighted disjunctions (i.e. fuzzy sets) of labels pertaining to the same vocabulary. The request is of the form:  $R = \bigwedge_{i \in A(R)} T_i$ , with  $T_i = \bigvee_{j \in R(T_i)} \lambda_{ij}/t_{ij}$  where  $A(R)$  is the set of attributes involved in the query,  $R(T_i)$  is the set of the terms involved in  $R$  for attribute  $i$ , and  $\lambda_{ij}$  is the level of preference of using  $t_{ij}$  for describing the request. However, using compound values, one can define new concepts that are not in  $\mathcal{T}_i$  for describing imprecise or fuzzy queries.

The terms are related through “possibilistic ontologies”  $O_i$   $i=1,n$  [7]. Relations in  $O_i$ ’s are modelled by necessity and possibility degrees:  $\Pi(t_{ij}, t_{ik}) = \Pi(t_{ik}, t_{ij})$  assesses to what extent  $t_{ij}$  and  $t_{ik}$  can refer to the same thing.  $N(t_{ij}, t_{ik})$  assesses to what extent it is certain that  $t_{ik}$  is a specialization of  $t_{ij}$ . Some important properties can be deduced from the characteristic properties of possibility and necessity measures:  $\Pi(t_{ij}, t_{ij}) = N(t_{ij}, t_{ij}) = 1$  and  $\Pi(t_{ij}, t_{ik}) = \Pi(t_{ik}, t_{ij})$ ,  $\Pi(t_{ij}, t_{ik}) \geq N(t_{ij}, t_{ik})$  (specialization supposes that the meanings overlap),  $N(t_{ij}, t_{ik}) > 0 \Rightarrow \Pi(t_{ij}, t_{ik}) = 1$  (if it’s somewhat certain, it has to be fully possible). Ontology’s relations can be extended using these properties together with the two following ones:

$$N(t_{ij}, t_{ih}) \geq \min(N(t_{ij}, t_{ik}), N(t_{ik}, t_{ih})), \quad (1)$$

which expresses the transitivity of the specialisation, together with the “hybrid transitivity” [5]:

$$\Pi(t_{ij}, t_{ih}) \geq N(t_{ij}, t_{ik}) * \Pi(t_{ik}, t_{ih}). \quad (2)$$

with  $a * b = b$  if  $b > 1 - a$  and  $a * b = 0$  otherwise. Request evaluation consists in retrieving all

data  $T'$  such that  $\Pi(R, T')$  or  $N(R, T')$  are non zero, which is made by computing  $\Pi(R, T') = \min_{i=1,n} \pi_i$ ,  $N(R, T') = \min_{i=1,n} \nu_i$ , with  $\pi_i = \max_{j \in R(T_i)} \min(\lambda_{ij}, \Pi(t_{ij}, t'_{ik}))$ ,  $\nu_i = \max_{j \in R(T_i)} \min(\lambda_{ij}, N(t_{ij}, t'_{ik}))$ , for possibility and necessity respectively, where  $T'_i = \{t'_{ik}\}$  for each  $i$ . Results are sorted first according to the decreasing values of  $N(R, T')$  and then according to the decreasing values of  $\Pi(R, T')$  for  $T'$  sharing the same value of  $N(R, T')$  [10].

Moreover, a weight  $\omega_i$  can be added to each fuzzy set  $T_i$  of terms pertaining to attribute  $i$  in the request, in order to express the relative importance of each attribute in the query. This leads to [6]:

$$\Pi(R, T') = \min_{i \in R(i)} \max(1 - \omega_i, \pi_i),$$

$$N(R, T') = \min_{i \in R(i)} \max(1 - \omega_i, \nu_i).$$

Lastly assume data are imprecise. Attributes values are then disjunctions of labels weighted with priority degrees. So the attribute value  $T'_i$  is a fuzzy set of labels:  $T'_i = \bigvee_{j \in D(T'_i)} \lambda'_{ij}/t'_{ij}$  where  $D(T'_i)$  is the set of terms involved in the attribute value, and  $\lambda'_{ij} \in [0, 1]$  is the term’s weight. We then have to use extended expressions [10]:

$$\pi_i = \max_{j \in R(T_i)} \min(\lambda_{ij}, \min_{k \in D(T'_i)} (\lambda'_{ik}, \Pi(t_{ij}, t'_{ik}))),$$

$$\nu_i = \max_{j \in R(T_i)} \min(\lambda_{ij}, \min_{k \in D(T'_i)} n_{ik})$$

$$\text{where } n_{ik} = \max(1 - \lambda'_{ik}, N(t_{ij}, t'_{ik})).$$

## 4 Multilingual Information Retrieval

Applying the ideas of section 3 to multilingual information retrieval (IR) systems raises some non-trivial issues. We first restate what multilingual IR is about, before defining a multilingual ontology and its use in a symbolic pattern matching procedure. The purpose of an IR system is to retrieve relevant documents in a collection. Relevance is defined according to a user query, typically a list of keywords, may be weighted, and aggregated using operators like *and* and *or*. Documents are stored as weighted lists of their significant words. The weight of a term  $t_i$  is estimated by combining the term frequency in the document, that is the number  $tf_{ij}$  of occurrences of  $t_i$  in document  $D_j$ , and the inverse document frequency of the term:  $idf_i = \log(d/df_i)$ , where  $df_i$  is the number of documents containing  $t_i$  and  $d$  is the total number of documents.  $idf_i$  can be considered as the entropy of  $t_i$ , that is the informa-

tion it gives. A document  $D_j$  is thus represented by:  $D_j = \{\rho_{ij}/t_i, i = 1, n\}$  where  $n$  is the total number of terms in the ontology and  $\rho_{ij}$  is the weight of the term  $t_i$  in document  $D_j$ , computed from  $tf_{ij}$  and  $idf_i$ , often as their product [8]. In multilingual IR system, documents are in different languages and whatever the query’s language, the system has to retrieve the relevant documents.

### 4.1 Multilingual Ontology

As just seen, and unlike in section 3, data representing documents are lists of weighted keywords. As a consequence, we have only one ontology, since there is only one domain for keywords. The ontology is used to define a controlled vocabulary which gathers valid terms for document indexing and for expressing requests. In the following, we will suppose that every term in the query and in the document index are in the ontology. By multilingual ontology we mean the following, based on EuroWordNet [11]. A “synset” is a set of synonymous labels, that is a clique of terms such as :  $S_h = \{t_{hi} \in \mathcal{T}\}$  with  $\forall(i, j), t_{hi} \neq t_{hj}$ ,  $\Pi(t_{hi}, t_{hj}) = 1$  and  $N(t_{hi}, t_{hj}) = N(t_{hj}, t_{hi}) = 1$ . i.e. the terms in a synset are supposed to be perfect synonyms.

Each term belongs to only one synset and w.r.t. (1) and (2) is synonymous of any other term in the synset, considering that a term  $t_{ij}$  is characterised by its meaning, and not only its label (in case of polysemic terms). Possibility and necessity relations for assessing approximate synonymy and specialisation are defined between synsets, as between the nodes of an ontology in section 3. A multilingual ontology is composed of a set of ontologies in the different languages. Synsets of different ontologies are related to each other with necessity and possibility degrees equal to 1 for modeling equivalences of terms between the language (Fig. 1). As in the monolingual ontology, the inter-lingual relations can be expanded using (1) and (2). Since these properties are language independent, we can deduce the matching of a query in some language with a document in any other language. However, both ontologies in two languages can have a different architecture, as in Fig. 1. Indeed, in this example, synset  $S_a$  can be translated into  $S'_a$ , and  $S_c$  into  $S'_c$ . But  $S_b$  has no translation defined in

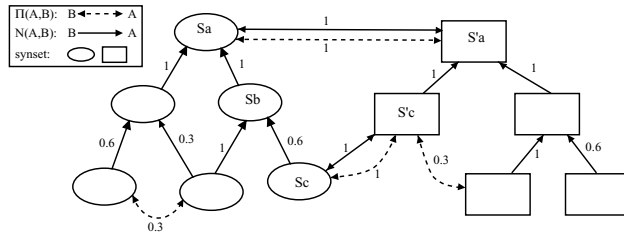


Figure 1: Multilingual ontology

the ontology. However, using transitivity property, the possibility and necessity relations between  $S_b$  and  $S'_a$  can be evaluated. Namely, we have  $N(S_a, S_b) = 1$  and  $N(S'_a, S_a) = 1$ . We can deduce that  $N(S'_a, S_b) = 1$ , but we have no information about  $N(S_b, S'_a)$ .

### 4.2 Possibilistic Indexing

To apply the symbolic pattern matching as defined in section 3, we have to estimate the relevance of the document, using possibility and necessity degrees. If these degrees can be estimated w.r.t. the terms in the document, using the ontology, request evaluation can be performed.

To be homogenous with the ontology repre-

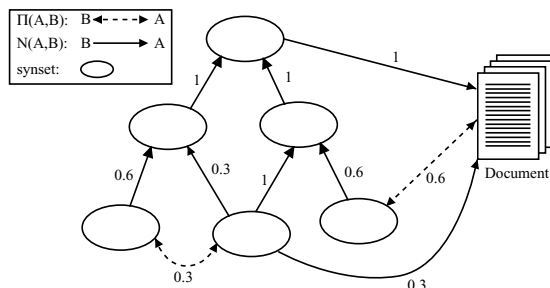


Figure 2: Document integrated in the ontology using possibility an necessity

sentation, a possibility and necessity degree of matching between the document and the synsets must be evaluated taking into account the weights of the terms in the document. Let us consider each document as a fuzzy set (e.g. [3, 9]). A weight  $\rho_{hij}$  of a term  $t_{hi}$  in a synset  $S_h$  w.r.t. document  $D_j$  is thus the grade of compatibility between  $t_{hi}$  and  $D_j$ :  $\rho_{hij} = \mu_{D_j}(t_{hi})$ . Given a synset  $S_h = \{t_{hi}, i = 1, p\}$ , we want to estimate to what extent the synset describes the document

$D_j$ , i.e.  $\Pi(S_h, D_j)$  and  $N(S_h, D_j)$ . As the terms in the synset are synonymous, we assume that each of them can describe the document as well. Note that in classical IR systems, synonyms are often aggregated with an *or* operator to expand the query. We have  $\Pi(S_h, D_j) = \max_i(\Pi(t_{hi}, D_j))$  and  $N(S_h, D_j) = \max_i(N(t_{hi}, D_j))$ . See Fig.3 . Considering that the weight  $\rho_{hij}$  is an intermedi-

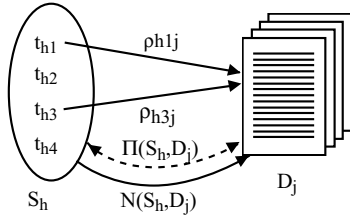


Figure 3: Linking a document to the ontology

ary degree between possibility and necessity that the term describes the document, the possibility and the necessity degrees will be estimated as [12]:

if  $\rho_{hij} < \frac{1}{2}$ ,  $\Pi(t_{hi}, D_j) = 2\rho_{hij}$  and  $N(t_{hi}, D_j) = 0$   
if  $\rho_{hij} \geq \frac{1}{2}$ ,  $\Pi(t_{hi}, D_j) = 1$  and  $N(t_{hi}, D_j) = 2\rho_{hij} - 1$ .

Fig.4 shows an example of partial ontology between French and English. Documents are

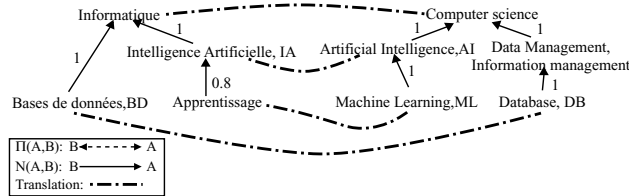


Figure 4: Example ontology

indexed using this ontology and statistical measures. Let’s consider an English document  $D$ , with the index shown in Table 1. This suggests that this document deals with

Table 1: Document Index

Term	$\rho$	$\Pi$	$N$
Computer Science	0	0	0
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

artificial intelligence, more specially with ma-

chine learning, applied to databases. Notice that despite “artificial intelligence” and “AI” have exactly the same meaning (they are in the same synset), their weights are different, since from a statistical point of view, the term “AI” is more frequent than “artificial intelligence” (the abbreviation is widely used instead of the full term). Thus, the  $(\Pi, N)$  degrees with the synset  $\{ArtificialIntelligence, AI\}$  is  $(\max(0.4, 1), \max(0, 0.4)) = (1, 0.4)$ . The *Computer Science* degree is 0 due to statistical reasons, even if the document subject deals with computer science.

### 4.3 Query Evaluation

Evaluating a query means estimating to what extent the expression of the query provides a good description for the document, or at least a part of it. Let’s use the above example to illustrate the evaluation, considering the query in French:  $R = BD \wedge IntelligenceArtificielle$  that is  $DB \wedge AI$  in English. We have  $\Pi(BD, Database) = N(BD, Database) = 1$  since they belong to synsets in one-to-one correspondence. Using (1) and (2), we can say that  $N(BD, D) = 0.2$  and  $\Pi(BD, D) = 1$ . In the same way,  $\Pi(IntelligenceArtificielle, AI) = N(IntelligenceArtificielle, AI) = 1$ . But to match  $D$ , we have to consider both the synset  $\{ArtificialIntelligence, AI\}$  and the singleton  $\{Machinelearning\}$ . Indeed, even if the term “Artificial Intelligence” is less frequent than “Machine learning”, we know that *Machine learning* (the concept) IS in *Artificial Intelligence*, thus the document  $D$  can be relevant. Evaluating  $\{ArtificialIntelligence, AI\}$  is obvious since it is the same case as  $BD$  and  $Database$  (i.e  $(\Pi, N) = (1, 0.4)$ ), but for  $\{Machinelearning\}$  there is two ways to evaluate possibility and necessity values, using transitivity:  $Intelligence Artificielle \rightarrow AI \rightarrow Machine Learning$  gives:  $N(IA, ML) = 1$  and  $\Pi(IA, ML) = 1$ .

$Intelligence Artificielle \rightarrow Apprentissage \rightarrow Machine Learning$  gives:  $\Pi(IA, ML) = 1$  and  $N(IA, ML) \geq 0.8$ . In addition,  $\Pi(ML, D) = 1$  and  $N(ML, D) = 0.6$ . This gives us degrees using *Machine Learning* path:  $\Pi'(IA, D) = 1$  and  $N'(IA, D) = 0.6$ . Since both values (direct

IA, and using expansion through *Machine Learning*) are possible, they are considered as disjunctive (as it is usual in IR systems) and we keep the max values between (1, 0.6) and (1, 0.4), thus  $\Pi(IA, D) = 1$  and  $N(IA, D) = 0.6$ .

We have supposed here that more specific documents (i.e containing more specific terms) are also relevant. If the user only wants to retrieve general documents, it should be possible to disable specific expansions, or to weight expanded results to reflect user's preferences. Note that only the score degrees from *Machine Learning* influence the result, as if the query were  $BD \wedge ML$  (for this document), since ML is actually IA. Identically, if the query were *Informatique*, the score wouldn't have been null, expanding the query.

The final query score will be:  $\Pi(R, D) = \min(\Pi(BD, D), \Pi(IA, D)) = 1$  and  $N(R, D) = 0.2$ . Several documents can then be sorted like in [10]. This simple example can be expanded using importance weights, as in section 3.

## 5 Concluding remarks

This paper is preliminary in many respects. Issues to be developed include: i) the assessment of the necessity and possibility degrees, especially in the ontology of each language, ii) the handling of the importance of keywords in the request, iii) the integration of classical FPM techniques for dealing with attributes such as date, size of documents, etc., iv) the modeling of the genuine focus of the request excluding documents which are too general or too specific. Moreover, the ideas presented here need to be tested in an implementation.

## Acknowledgements

This work has been supported by the E-Court European project (IST-2000-28199).

## References

[1] T. Andreasen, H. Christiansen, and H. L. Larsen. *Flexible Query Answering Systems*. Kluwer Academic Publishers, 1997.

[2] A. Bidault, C. Froidevaux, and B. Safar. Proximité entre requêtes dans un contexte médiateur. *RFIA 2002*, 2:653–662, 2002.

[3] D.A. Buell. An analysis of some fuzzy subset applications to information retrieval systems. *Fuzzy Sets and Systems*, 7(1):35–42, 1982.

[4] M. Cayrol, H. Farreny, and H. Prade. Fuzzy pattern matching. *Kybern.*, 11:103–16, 1982.

[5] D. Dubois and H. Prade. Resolution principles in possibilistic logic. *Int. Jour. of Approximate Reasoning*, 4(1):1–21, 1990.

[6] D. Dubois, H. Prade, and C. Testemale. Weighted fuzzy pattern matching. *Fuzzy Sets and Systems*, 28:313–331, 1988.

[7] H. Farreny and H. Prade. Dealing with vagueness of natural languages in man-machine communication. In W. Karwowski and A. Mital, editors, *Applications of Fuzzy Set Theory in Human Factors*, pages 71–85. Elsevier, 1986.

[8] D.A. Grossman and O. Frieder. *Information Retrieval: Algorithms and Heuristics*. Kluwer Academic Publishers, 1998.

[9] D. Kraft, G. Bordogna, and G. Pasi. Fuzzy set techniques in information retrieval. In J. Bezdek et al., editor, *Fuzzy Sets in Approximate Reasoning and Information Systems*, pages 469–510. Kluwer, 1999.

[10] Y. Loiseau and H. Prade. Qualitative pattern matching with linguistic terms. pages 125–134. STarting AI Res.Symp., Lyon, IOS Press, July 2002.

[11] University of Amsterdam. Eurowordnet. <http://www.hum.uva.nl/~ewn/>.

[12] H. Prade and C. Testemale. Application of possibility and necessity measures to documentary information retrieval. *LNCS*, 286:265–275, 1987.

[13] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problem of ambiguity in natural language. *J. of Art. Int. Res.*, 1998.

[14] L.A. Zadeh. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, 4(2):103–111, 1996.