A novel probabilistic quantifier fuzzification mechanism for information retrieval

Félix Díaz-Hermida David E. Losada Alberto Bugarín Senén Barro
Intelligent Systems Group
Department of Electronics and Computer Science
University of Santiago de Compostela
15782 Santiago de Compostela
{felixdh,dlosada,alberto,senen}@dec.usc.es

Abstract

In this work, a novel quantifier fuzzification mechanism is proposed. This method is deeply rooted in the theory of probability and skips the nested assumption for crisp representatives, which is often taken by other probabilistic approaches to quantification. The new proposal takes into account all possible crisp representatives which yields to a natural and intuive strategy for information retrieval tasks. Furthermore, preliminar analysis of the formal properties of the fuzzification mechanism permits us to advance that the application of this method in other domains is also promising.

1 Introduction

Fuzzy quantifiers have been extensively applied in diverse fields such as expert systems, monitoring and control of processes, database systems, etc. [2, 12]. These fuzzy tools have played a key role in such domains because the linguistic statements a human expert uses are naturally modeled and, hence, the expressiveness of the system is enriched.

Fuzzy quantifiers can be defined in a direct way, e.g. proposing a form of combination of the membership values of the elements belonging to the involved fuzzy set(s). Nevertheless, given a certain linguistic expression, it is often difficult to achieve consensus about the most appropriate quantified definition. In order to avoid such inconveniences, indirect definitions of fuzzy quantifiers have been introduced [9], based on the concept of semi-fuzzy quantifier, which is a halfway point between classic quantifiers and fuzzy quantifiers. Fuzzy quantifiers are intuitively defined from semi-fuzzy quantifiers through a so-called quantifier fuzzification mechanism. Because semi-fuzzy quantifiers are closer to the well-known classic quantifiers, the implementation of a linguistic expression in terms of semi-fuzzy quantifiers results more natural and intuitive.

In the information retrieval (IR) literature, fuzzy quantification has been applied for designing flexible query languages [1]. Since the connection between query term semantics and document contents is inherently vague, the retrieval process can be naturally modelled in terms of fuzzy sets. Fuzzy quantification supplies appropriate formal tools for handling linguistic expressions which enrich the query languages of the IR system. This aids users to establish additional constraints in the retrieval process (e.g. retrieved documents should match at least 3 of the query terms).

The importance of fuzzy quantifiers for IR was empirically demonstrated for large collections of documents in [11]. Nevertheless, this practical deployment of fuzzy quantifiers also revealed that a new class of quantifier fuzzification mechanisms may be beneficial. This motivated us to define here a new fuzzification method which is evaluated for a retrieval task. Furthermore, the most relevant properties the model fulfills are presented, advancing the adequacy of this new approach for other domains.

The rest of this paper is organized as follows. Sec-

tion 2 reports briefly some background concepts on fuzzy quantification. Section 3 explains the new quantification proposal and section 4 applies the new quantification framework for handling information retrieval. The paper ends with some conclusions.

2 Fuzzy quantifiers

The formal notions of classic quantifier, fuzzy quantifier and semi-fuzzy quantifier have been used to associate meaning to quantified sentences [9]. Formally, a classic s-ary quantifier on a base or referential set E is a mapping $Q : \wp(E)^s \longrightarrow \{0,1\}$, where $\wp(E)$ is the powerset of E. Throughout this work we assume the referential set E to be finite, which is sufficient from a practical perspective.

An s-ary fuzzy quantifier $\widetilde{Q}$ on a base set $E \neq \varnothing$ is a mapping $\widetilde{Q} : \widetilde{\wp}(E)^s \longrightarrow [0,1]$ which to each choice of $X_1, \ldots, X_s \in \widetilde{\wp}(E)$ assigns a gradual result $\widetilde{Q}(X_1, \ldots, X_s) \in [0,1]$ ($\widetilde{\wp}(E)$ is the fuzzy powerset of E).

In many cases, it is not easy to achieve consensus on an intuitive and generally applicable expression for implementing a given quantified sentence. To overcome this problem, the concept of semi-fuzzy quantifier was introduced [9]. A semi-fuzzy quantifier is a half-way point between classic quantifiers and fuzzy quantifiers, which is very close to the idea of Zadeh's linguistic quantifier [14]. Semi-fuzzy quantifiers are similar to classic quantifiers, but they allow variation of the results in $[0,1]$. Formally, an s-ary semi-fuzzy quantifier Q on a base set $E \neq \varnothing$ is a mapping $Q : \wp(E)^s \longrightarrow [0,1]$ which assigns a gradual result $Q(X_1, \ldots, X_s) \in [0,1]$ to each choice of crisp $X_1, \ldots, X_s \in \wp(E)$.

Semi-fuzzy quantifiers are much more intuitive and easier to define than fuzzy quantifiers, but they do not resolve the problem of evaluating fuzzy quantified sentences. In order to do so quantifier fuzzification mechanisms are needed [9] that enable us to transform semi-fuzzy quantifiers into fuzzy quantifiers, i.e. mappings with domain in the universe of semi-fuzzy quantifiers and range in the universe of fuzzy quantifiers:

$$F : (Q : \wp(E)^s \to [0,1]) \to \left(\widetilde{Q} : \widetilde{\wp}(E)^s \to [0,1]\right)$$

2.1 Quantifier fuzzification mechanisms

Several methods for evaluating quantified sentences have been proposed in the literature [9, 5]. In [5] two models for the evaluation of fuzzy quantified sentences are proposed. These models show a very consistent behaviour and are based on a voting model interpretation of fuzzy sets [10, 6]. If the universe of discourse E is finite and expressions are unary (i.e. involve a single fuzzy set) then both models collapse into the same:

$$\widetilde{Q}(X) = (F(Q))(X) = \sum_{i=0}^{m} Q\left((X)_{\geq \alpha_i}\right) \cdot (\alpha_i - \alpha_{i+1}) \quad (1)$$

where $\alpha_0 = 1, \alpha_{m+1} = 0$ and $\alpha_1 \geq \ldots \geq \alpha_m$ denote the membership values in descending order of the elements in E to the fuzzy set X and $(X)_{\geq \alpha_i}$ stands for the α-cut of level α_i of X, i.e. the crisp set containing the elements of E whose degree of membership in X is greater or equal than α_i. For the unary and binary cases, expression 1 is equivalent to the quantification model defined in [3][1]. Moreover, for the case of non-decreasing unary quantifiers, it is equivalent to the quantification method based on ordered weighted operators [13].

In equation 1, the value $\alpha_i - \alpha_{i+1}$ can be interpreted as the probability that $(X)_{\geq \alpha_i}$ is selected as the crisp representative for the fuzzy set X. Therefore, the semi-fuzzy quantifier is applied for every crisp representative of X and those values are weighted by the probability of each crisp representative. In this formulation, the use of α-cuts makes that, given the fuzzy set X, the crisp representatives $(X)_{\geq \alpha_i}$ are nested.

In [11], equation 1 was empirically evaluated for the basic IR task. Although this experimentation made evident the benefits that IR might obtain from fuzzy quantifiers, it also revealed that the nested assumption may not always be appropriate, as it will be seen in the following sections. This motivated us to propose a novel probabilistic

[1] keeping aside differences related to representative normalization.

method that skips the nested assumption. In section 4 we will enter into details on the adequacy of the new quantification proposal for IR. Along this paper, the approach sketched in equation 1 will be referred to as NVM (standing for Nested Voting Model) approach.

3 A probabilistic interpretation of fuzzy sets

Given a fuzzy set $X \in \widetilde{\wp}(E)$, the process that selects a number of elements in E to belong to a crisp representative of X can be viewed as a random process in which n *mutually independent* binary decisions are made ($n = |E|$). Every individual decision involving an element $e \in E$ may be viewed as a Bernoulli trial whose probability of success (i.e. the probability of selecting e for representing X) is equal to $\mu_X(e)$. Hence, for every possible crisp representative of X, $Y \in \wp(E)$, we can estimate its probability as follows. Given a discrete random variable $Representative_X$ which takes values on $\wp(E)$, the probability that $Representative_X$ results in Y is equal to:

$$P(Representative_X = Y) = \prod_{e \in Y} \mu_X(e) \cdot \prod_{e \notin Y} (1 - \mu_X(e))$$

For simplicity, we introduce the following compact notation: $m_X(Y) = P(Representative_X = Y)$.

In the next section this definition is used for designing a novel quantifier fuzzification mechanism based on this independence assumption.

3.1 A new fuzzification mechanism

Following the previous definition, a new fuzzification method in which all possible crisp representatives of a given fuzzy set X are considered arises in a natural way. This contrasts with the NVM approach in which only the nested crisp sets obtained by successive α-cuts on X are taken into account.

Definition 1 (F^A) *Let $Q : \wp(E)^s \to [0,1]$ be a semi-fuzzy quantifier. We define the quantifier fuzzification mechanism F^A as:*

$$F^A(Q)(X_1, \ldots, X_s) = \sum_{Y_1 \in \wp(E)} \cdots \sum_{Y_s \in \wp(E)} m_{X_1}(Y_1) \ldots m_{X_s}(Y_s)\, Q(Y_1, \ldots, Y_s)$$
$$X_1, \ldots, X_s \in \widetilde{\wp}(E)$$

Following this definition, all the crisp representatives are handled independently and no crisp representative is disregarded a priori. The superindex A was chosen to stress that all crisp representatives are considered.

Unfortunately, in the general case F^A is not computable in polynomial time. Nevertheless, when *quantitative* semi-fuzzy quantifiers (i.e, those which can be expressed as a function of the cardinalities of the involved sets[2]) are handled, it is possible to develop polynomial time algorithms. This is very important because quantitative quantifiers are the most interesting from a practical view [4, 7] and, indeed, sufficient for our purposes in IR.

Now we will sketch the procedure for solving quantitative unary quantifiers. Algorithms for solving higher arity quantitative quantifiers can be designed using similar ideas.

We will denote by $E^i = \{e_1, \ldots, e_i\}$ a referential containing i elements. By $Y^i \in \wp\left(E^i\right)$ ($X^i \in \widetilde{\wp}\left(E^i\right)$) we will denote a crisp (fuzzy) set on this referential.

Let us consider a unary semi-fuzzy quantitative quantifier:

$$Q_1\left(Y^i\right) = q_1\left(\left|Y^i\right|\right), X^i \in \wp\left(E^i\right)$$

where q_1 is a function with the form $q_1 : \mathbb{N} \to [0,1]$.

For this case the independence quantification expression becomes:

[2] More specifically, those which can be expressed as a function of the cardinalities of the arguments and their boolean combinations[8].

$$
\begin{aligned}
F^A(Q_1)\left(X^i\right) &= \sum_{Y^i \in \wp(E^i)} m_{X^i}\left(Y^i\right) Q_1\left(Y^i\right) = \\
&= \sum_{Y^i \in \wp(E^i)||Y^i|=0} m_{X^i}\left(Y^i\right) Q_1\left(Y^i\right) + \ldots \\
&+ \sum_{Y^i \in \wp(E^i)||Y^i|=i} m_{X^i}\left(Y^i\right) Q_1\left(Y^i\right) = \\
&= \sum_{Y^i \in \wp(E^i)||Y^i|=0} m_{X^i}\left(Y^i\right) q_1(0) + \ldots \\
&+ \sum_{Y^i \in \wp(E^i)||Y^i|=i} m_{X^i}\left(Y^i\right) q_1(i)
\end{aligned}
$$

If we denote $\sum_{Y^i \in \wp(E^i)||Y^i|=j} m_{X^i}\left(Y^i\right)$ by $\Pr\left(card_{X^i}=j\right)$[3] we can rewrite the previous expression as follows:

$$
\begin{aligned}
F^A(Q_1)\left(X^i\right) &= \Pr\left(card_{X^i}=0\right) q_1(0) + \ldots + \\
&+ \Pr\left(card_{X^i}=i\right) q_1(i) = \\
&= \sum_{j=0}^{i} \Pr\left(card_{X^i}=j\right) q_1(j)
\end{aligned}
$$

It can be proved that the values $\Pr\left(card_{X^i}=j\right)$ can be obtained with a complexity $\mathcal{O}(n^2)$.

The next example clarifies the use of the new quantification approach. First, probabilities of all possible crisp representatives are computed and, next, the previous expression is applied.

Example 1 *Let us consider the evaluation of the quantified sentence "almost all students are tall". Suppose that we model the property tall for a number of individuals $E=\{e_1,e_2,e_3\}$ through the fuzzy set $tall=\{0.8/e_1, 0.9/e_2, 1/e_3\}$ and we support the quantified expression "almost all" by means of the following semi-fuzzy quantifier:*

$$
\begin{aligned}
Q(X) &= q_1(|X|) \\
q_1(n) &= \left(\frac{n}{3}\right)^2
\end{aligned}
$$

First, we compute the probabilities $\Pr\left(card_{tall}=j\right)$ for every value of j:

$$
\begin{aligned}
\Pr\left(card_{tall}=0\right) &= \sum_{Y^i \in \wp(E)||Y^i|=0} m_{tall}\left(Y^i\right) = \\
&= m_{tall}(\emptyset) = 0.2 \cdot 0.1 \cdot 0 = 0 \\
\Pr\left(card_{tall}=1\right) &= \sum_{Y^i \in \wp(E)||Y^i|=1} m_{tall}\left(Y^i\right) = \\
&= m_{tall}(\{e_1\}) + m_{tall}(\{e_2\}) + \\
&+ m_{tall}(\{e_3\}) = 0.8 \cdot 0.1 \cdot 0 + \\
&+ 0.2 \cdot 0.9 \cdot 0 + 0.2 \cdot 0.1 \cdot 1 = 0.02 \\
\Pr\left(card_{tall}=2\right) &= \sum_{Y^i \in \wp(E)||Y^i|=2} m_{tall}\left(Y^i\right) = \\
&= m_{tall}(\{e_1,e_2\}) + m_{tall}(\{e_1,e_3\}) + \\
&+ m_{tall}(\{e_2,e_3\}) = 0.8 \cdot 0.9 \cdot 0 + \\
&+ 0.8 \cdot 0.1 \cdot 1 + 0.2 \cdot 0.9 \cdot 1 = \\
&= 0.08 + 0.18 = 0.26 \\
\Pr\left(card_{tall}=3\right) &= \sum_{Y^i \in \wp(E)||Y^i|=3} m_{tall}\left(Y^i\right) = \\
&= m_{tall}(\{e_1,e_2,e_3\}) = \\
&= 0.8 \cdot 0.9 \cdot 1 = 0.72
\end{aligned}
$$

And then,

$$
\begin{aligned}
F^A(Q)(tall) &= \sum_{j=0}^{3} \Pr\left(card_{tall}=j\right) q_1(j) \\
&= \Pr\left(card_{tall}=0\right) q_1(0) + \\
&+ \Pr\left(card_{tall}=1\right) q_1(1) + \\
&+ \Pr\left(card_{tall}=2\right) q_1(2) + \\
&+ \Pr\left(card_{tall}=3\right) q_1(3) = \\
&= 0 \cdot 0 + 0.02 \cdot \left(\frac{1}{3}\right)^2 + 0.26 \cdot \left(\frac{2}{3}\right)^2 + \\
&+ 0.72 \cdot \left(\frac{3}{3}\right)^2 = 0.838
\end{aligned}
$$

It is interesting to see how the NVM approach proceeds against the same example. Given the fuzzy set $tall$, the α_i values (equation 1) are $\alpha_0=1, \alpha_1=1, \alpha_2=0.9, \alpha_3=0.8, \alpha_4=0$ and the fuzzification process runs as follows:

$$
\begin{aligned}
(F(Q))(tall) &= \sum_{i=0}^{3} Q\left((tall)_{\geq \alpha_i}\right) \cdot (\alpha_i - \alpha_{i+1}) \\
&= Q\left((tall)_{\geq \alpha_0}\right) \cdot (\alpha_0 - \alpha_1) + \\
&+ Q\left((tall)_{\geq \alpha_1}\right) \cdot (\alpha_1 - \alpha_2) + \\
&+ Q\left((tall)_{\geq \alpha_2}\right) \cdot (\alpha_2 - \alpha_3) + \\
&+ Q\left((tall)_{\geq \alpha_3}\right) \cdot (\alpha_3 - \alpha_4) = \\
&= Q(\{e_3\}) \cdot (1-1) + Q(\{e_3\}) \cdot (1-0.9) + \\
&+ Q(\{e_2,e_3\}) \cdot (0.9-0.8) + \\
&+ Q(\{e_1,e_2,e_3\}) \cdot (0.8-0) = \\
&= \left(\frac{1}{3}\right)^2 \cdot 0.1 + \left(\frac{2}{3}\right)^2 \cdot 0.1 + \left(\frac{3}{3}\right)^2 \cdot 0.8 \\
&= 0.855
\end{aligned}
$$

This example clarifies the differences between both methods. For instance, the NVM approach

[3] This value can be interpreted as the probability that the fuzzy set X^i is represented by a crisp set whose size is j.

estimates the odds that exactly two individuals are tall by means of the two elements of $tall$ having the largest degrees of membership, e_2, e_3. It is implicitly assumed that, if only two individuals are considered tall, these should be e_2 and e_3 mandatorily. This assumption is completely reasonable if the fuzzy set $tall$ was built from measures of height of the individuals involved. In that case, it is difficult to imagine a situation in which, for instance, e_3 and e_1 are considered as tall individuals whereas e_2 is not considered a tall person. Nevertheless, think that the fuzzy set $tall$ may represent predictions about the height of future descendants for three couples (e.g. estimated from the height of the members of each couple) and, hence, it might be the case that e_3 and e_1 do finally produce tall descendants whereas e_2 produces a short one. Summing up, it may be adequate to consider all possible two-sized crisp representatives for computing the odds that exactly two elements do comply with the property formalized by the fuzzy set. This is precisely what the F^A method does. The odds that exactly two individuals are tall are computed taking into account the odds that only e_1 and e_2 are tall, the odds that only e_2 and e_3 are tall and the odds that only e_1 and e_3 are tall.

As we will detail later, IR is foundated on a number of useful heuristics that have played a fundamental role to enhance retrieval performance, e.g. tf/idf to weight the importance of a term for a given document. Of course, this class of heuristics is not perfect and, hence, one can never be sure that the two terms which are the most significative in the context of a given document are those ones have the higher tf/idf values. As a consequence, the F^A approach is a good support for our application of fuzzy quantifiers in IR.

3.2 Properties of the model

A formal analysis of the properties fulfilled by the new fuzzification approach is currently undergoing. In this study we follow the axiomatic framework presented in [9]. We can advance that the model is well-behaved because it fulfills the properties of *correct generalization of crisp expressions, induced operations, external negation, internal negation, duality, internal meets, monotonicity in arguments, monotonicity in quantifiers and coherence with logic.* This assures that the quantification method proposed yields a natural and intuitive modeling of quantified expressions, as depicted briefly in the following examples. For instance, if external negation is not fulfilled sentences such as "at most 10 tall individuals are blonde" and "not more than 10 tall individuals are blonde" are not considered equivalent. The sentences "all tall individuals are blonde" and "no tall individual is not blonde" are only equivalent when the quantification model complies with internal negation. Duality assures that "some tall individuals are blonde" and "not all tall individuals are not blonde" are equivalent and the equivalence between "some tall individuals are blonde" and "there is some individual who is tall and blonde" is guaranteed by the property of internal meets. Monotonicity in quantifiers assures that the result of evaluating an expression such as "about 80% or more of the tall individuals are blonde" is less or equal than the result obtained from "about 60% or more of the tall individuals are blonde". These examples show clearly that unacceptable and counterintuitive situations might arise when the quantification approach does not comply with some of these fundamental properties. Since the F^A method complies with such properties, its application in a wide range of domains is promising.

4 Application in information retrieval

The adequacy of fuzzy quantifiers for IR was already anticipated in [1]. In a recent work [11], a query language enriched with quantified statements was empirically tested. This evaluation revealed that fuzzy quantifiers are beneficial in terms of retrieval performance. The proposal enclosed in [11] designs a general framework based on the NVM method in which quantifiers with different degrees of expressiveness can be handled. In the experimental setting quantified expressions were handled through unary quantifiers. This approach subsumes the quantification model based on ordered weighted operators [13] and, as argued before, it falls into the nested assumption for crisp representatives.

Next paragraphs sketch the use and convenience of the F^A model in the context of the basic IR task.

Consider a query formulation with the form

$at\_least\_3(qt_1, \cdots, qt_n)$. Given a document d_k of the document base, every query term qt_i produces a score which represents the connection between the document's semantics and the term. The set of n scores is then combined applying the quantifier. Formally, every document d_k induces a fuzzy set C_{d_k} on the set of query terms which is defined applying the popular tf/idf weighting strategy:

$$C_{d_k} = \{\mu_{C_{d_k}}(qt_1)/qt_1, \cdots, \mu_{C_{d_k}}(qt_n)/qt_n\}$$
$$\mu_{C_{d_k}}(qt_i) = \frac{f_{qt_i,k}}{max_z f_{z,k}} * \frac{idf(qt_i)}{max_l idf(qt_l)}$$

where $f_{qt_i,k}$ is the raw frequency of term qt_i in the document d_k and $max_z f_{z,k}$ is the maximum raw frequency computed over all terms mentioned by the document d_k. By $idf(qt_i)$ we refer to a function computing an inverse document frequency (idf) factor. For instance, it might be defined as $idf(qt_i) = log(max_l n_l/n_i)$, where n_i is the number of documents in which the term qt_i appears and the maximum $max_l n_l$ is computed over all terms in the indexing vocabulary. The value $idf(qt_i)$ is divided by $max_l idf(qt_l)$, which is the maximum value of the function idf computed over all terms in the alphabet. The idf factor tries to capture how significative is a term taking into account its global frequency in the whole collection. A term which is very frequent in the collection is not a good candidate for discriminating between relevant and unrelevant documents and, thus, it receives a low idf value. On the contrary, very infrequent terms are assigned high idf values because they are potentially good discriminators between relevance and non-relevance. Putting all together, the final tf/idf value makes that an ideal term is one that appears many times in the involved document (d_k) but very few times in the rest of the document collection.

The fuzzy set C_{d_k} models the connection between the document d_k and every query component. Quantification can now be applied on C_{d_k} for evaluating the quantified symbol at_least_3. Every quantification symbol Q will be supported by a given semi-fuzzy quantifier Q_s. The probabilistic fuzzification process is fired on Q_s yielding a fuzzy quantifier which is applied on C_{d_k}.

A key issue regards the selection of the semi-fuzzy quantifier associated to a given quantification symbol. For instance, a direct crisp implementation of the at_least_3 statement[4] is too strict for IR purposes because a document matching 8 query terms is considered as good as one matching 3 terms. As argued in [11] at least quantifiers can be good for retrieval purposes if implemented in a relaxed form. In particular, half-way implementations between a crisp atleast and a linear implementation[5] are promising:

$$\begin{aligned} PQ(X) &= pq(|X|) \\ pq(x) &= \frac{x^{exp}}{n^{exp}} \end{aligned}$$

where n is the number of query terms, which is used for obtaining an upper bound value for normalization purposes. It is not strange that non-relevant documents match a few query terms simply by chance. To minimize this problem the relaxed formulation makes that documents matching few terms receive a lower score compared to an alternative linear implementation. On the other hand, as the number of query-document matches grows, the contribution from every single match is more important. This reflects the intuition that an additional matching term is much more important if the document already matched a significant number of query terms.

Let us now suppose that we apply the previous power function (exp=2) for supporting the at_least_3 symbol[6]. Imagine a query $at\_least\_3(qt_1, qt_2, qt_3, qt_4)$ and consider a document d_k whose fuzzy set induced on the query components is $C_{d_k} = \{0.7/qt_1, 0.3/qt_2, 0/qt_3, 0.2/qt_4\}$. Applying now the fuzzification process explained along this paper, the query-document matching is assigned a score:

[4] which assigns 1 for sets whose cardinality is higher or equal than 3 and 0 otherwise.

[5] A linear implementation is common in popular IR matching functions, e.g. a simple addition of the query-document matching scores [11].

[6] This is only an example which aims at illustrating the quantification process. Of course, at_least_k statements will ideally be supported by semi-fuzzy quantifiers whose support functions will be dependent on k. For the sake of clarity this class of functions are skipped here and we refer to [11] for a report on the practical behaviour of different functions supporting at_least_k expressions.

$$\Pr\left(card_{C_{d_k}}=0\right)pq\left(0\right)+\Pr\left(card_{C_{d_k}}=1\right)pq\left(1\right)+$$
$$+\Pr\left(card_{C_{d_k}}=2\right)pq\left(2\right)+\Pr\left(card_{C_{d_k}}=3\right)pq\left(3\right)+$$
$$+\Pr\left(card_{C_{d_k}}=4\right)pq\left(4\right)=(0.3\cdot 0.7\cdot 0.8)\cdot 0+$$
$$+(0.7\cdot 0.7\cdot 0.8+0.3\cdot 0.3\cdot 0.8+0.3\cdot 0.7\cdot 0.2)\cdot \tfrac{1}{4^2}+$$
$$+(0.7\cdot 0.3\cdot 0.8+0.7\cdot 0.7\cdot 0.2+0.3\cdot 0.3\cdot 0.2)\cdot \tfrac{2^2}{4^2}+$$
$$+(0.7\cdot 0.3\cdot 0.2)\cdot \tfrac{3^2}{4^2}=0.12625$$

which reflects the fact that it is unlikely that the document's contents are actually related to at least three out of the four query terms. Note that, as argued in section 3, the values $\Pr\left(card_{C_{d_k}}=j\right)$ can be obtained in quadratic time and, hence, the matching function is polynomial. This assures an efficient retrieval process.

Let us now apply the NVM approach to handle the same example. The score assigned is equal to:

$$\sum_{i=0}^{4} PQ\left(\left(C_{d_k}\right)_{\geq\alpha_i}\right)\cdot(\alpha_i-\alpha_{i+1})$$

where $\alpha_0=1,\alpha_1=0.7,\alpha_2=0.3,\alpha_3=0.2,\alpha_4=0$ and $\alpha_5=0$. It follows that the final value yielded by the NVM method is:

$$PQ\left(\emptyset\right)\cdot 0.3+PQ\left(\{qt_1\}\right)\cdot 0.4+$$
$$+PQ\left(\{qt_1,qt_2\}\right)\cdot 0.1+$$
$$+PQ\left(\{qt_1,qt_2,qt_3\}\right)\cdot 0.2+$$
$$+PQ\left(\{qt_1,qt_2,qt_3,qt_4\}\right)\cdot 0=$$
$$=\tfrac{1}{4^2}\cdot 0.4+\tfrac{2^2}{4^2}\cdot 0.1+\tfrac{3^2}{4^2}\cdot 0.2=0.1625$$

Observe how NVM measures the odds that only one query term is related to d_k by means of the odds that qt_1 is related to d_k (i.e. all the other query terms are disregarded because qt_1 is the most probable term). On the contrary, the novel probabilistic quantification method computes all the possibilities (this is accounted by the value $(0.7\cdot 0.7\cdot 0.8+0.3\cdot 0.3\cdot 0.8+0.3\cdot 0.7\cdot 0.2)$). This is the most natural choice. Although qt_1 is the term whose connection with d_k contents is the most probable, it might be the case that d_k is not actually related to qt_1 but it is related to another query term whose degree of membership in C_{d_k} is lower. Think that query-document connection is measured by heuristics such as tf/idf which, of course, are not perfect. Hence, the fairest decision is to account for all possible cases.

Experiments on the effect of the new probabilistic fuzzification method on retrieval performance are currently undergoing. Preliminar tests against a subset of the TREC collection (composed of Wall Street Journal articles) allow us to advance that the novel quantification proposal yields better retrieval performance results. In terms of average precision, the new method is 6.5% better than the NVM approach. This is especially remarkable because the NVM mechanism was already able to overcome popular IR matching strategies [11].

5 Conclusions and future work

A novel quantifier fuzzification mechanism was proposed. This method is foundated on the theory of probability and skips the nested assumption on crisp representatives. The new quantification proposal handles query quantified statements in a natural and intuitive way. There is also sound theoretical evidence on its good behaviour and preliminar empirical evidence about its benefits. Although the design of the fuzzification approach was inspired by previous applications of fuzzy quantifiers for information retrieval, we are also confident about its use in other application domains. This belief is foundated on the relevant properties the model fulfills.

A number of research lines remain to be explored. We have still not studied the connection between quantification and relevance feedback. It might be the case that fuzzy quantifiers could also play a valuable role in the context of a feedback cycle. Also, document or term clustering methods might be assisted by fuzzy quantification using similar techniques to those proposed here. Moreover, more experimental effort is needed to compare our approach with other quantification methods proposed in the literature.

6 Acknowledgments

Authors wish to acknowledge support from the Spanish Ministry of Education and Culture through grants TIC2000-0873 and TIC2003-09400-C04-03. D. E. Losada is supported by the "Ramón y Cajal" R&D program, which is funded in part by "Ministerio de Ciencia y Tecnología" and in part by FEDER funds.

References

[1] G. Bordogna and G. Pasi. Linguistic aggregation operators of selection criteria in fuzzy information retrieval. *International Journal of Intelligent Systems*, 10(2):233–248, 1995.

[2] P. Bosc, L. Lietard, and O. Pivert. Quantified statements and database fuzzy querying. In P.Bosc and J. Kacprzyk, editors, *Fuzziness in Database Management Systems*, volume 5 of *Studies in Fuzziness*, pages 275–308. Physica-Verlag, 1995.

[3] M. Delgado, D. Sánchez, and M. A. Vila. Fuzzy cardinality based evaluation of quantified sentences. *International Journal of Approximate Reasoning*, 23(1):23–66, 2000.

[4] F. Díaz-Hermida, A. Bugarín, and S. Barro. Definition and classification of semi-fuzzy quantifiers for the evaluation of fuzzy quantified sentences. *International Journal of Approximate Reasoning*, 34(1):49–88, 2003.

[5] F. Díaz-Hermida, A. Bugarín, P. Cariñena, and S. Barro. Voting model based evaluation of fuzzy quantified sentences: a general framework. *Fuzzy Sets and Systems*. Accepted.

[6] D. Dubois and H. Prade, editors. *Possibility theory : An approach to computerized processing of uncertainty*. Plenum, 1988.

[7] I. Glöckner. Evaluation of quantified propositions in generalized models of fuzzy quantification. Technical report, Universität Bielefeld, January 2003. Preprint submitted to the International Journal on Approximate Reasoning, Elsevier Science, 15th January 2003.

[8] I. Glöckner. *Fuzzy Quantifiers in Natural Language: Semantics and Computational Models*. PhD thesis, Universität Bielefeld, 2003.

[9] I. Glöckner and A. Knoll. A formal theory of fuzzy natural language quantification and its role in granular computing. In W. Pedrycz, editor, *Granular computing: An emerging paradigm*, volume 70 of *Studies in Fuzziness and Soft Computing*, pages 215–256. Physica-Verlag, 2001.

[10] R. Kruse, J. Gebhardt, and F. Klawonn. *Foundations of Fuzzy Systems*. John Wiley and Sons Inc, 1994.

[11] D. E. Losada, F. Díaz-Hermida, A. Bugarín, and S. Barro. Experiments on using fuzzy quantified sentences in adhoc retrieval. In *Proc. SAC-04, the 19th ACM Symposium on Applied Computing - Special Track on Information Access and Retrieval*, Nicosia, Cyprus, March 2004.

[12] R. R. Yager. Approximate reasoning as a basis for rule-based expert systems. *IEEE Transactions on Systems, Man and Cybernetics*, 14(4):636–642, 1984.

[13] R.R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man and Cybernetics*, 18(1):183–191, 1988.

[14] L.A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Comp. and Machs. with Appls.*, 8:149–184, 1983.