

Linguistically quantified thresholding strategies for text categorization

Slawomir Zadrozny

Systems Research Institute
Polish Academy of Sciences.
Warszawa, Poland
zadrozny@ibspan.waw.pl

Janusz Kacprzyk

Systems Research Institute
Polish Academy of Sciences.
Warszawa, Poland
kacprzyk@ibspan.waw.pl

Abstract

A new thresholding strategy for a text categorization problem is proposed. It is based on Zadeh's calculus of linguistically quantified propositions. The strategy may be also interpreted in terms of fuzzy integral.

Keywords: text categorization, thresholding strategy, linguistic quantifiers.

1 Introduction

Text categorization is one of problems considered within *information retrieval (IR)* cf. [5-8]. Basically, it is an example of the classification problem that is meant as: having a set of textual documents of known categories of new, unlabelled documents has to be decided. Most interesting is a case where there are more than two categories (classes) and more than one category may be assigned to a document. This is so-called *multiclass multilabel* categorization problem. A whole array of machine learning algorithms is here applicable to create a suitable classifier. Usually, such a classifier yields a ranked list of categories possibly relevant for a document. The ranking is based on a certain score computed for each category and a document under consideration. Usually, such a score does not have a clear absolute interpretation [7]. Hence, it is not possible to unambiguously indicate what score is high enough to assign corresponding category to a document. This problem, referred to as *the thresholding strategy* [7,8] has been addressed in a few publications concerning text categorization. The very same problem may be also considered for the

results of queries in the context of both information retrieval and database querying [2]. For the former context, it is widely adopted that a query returns a set of matching documents additionally associated with a *return status value (RSV)* being a number from some fixed interval (often $[0,1]$). Such a number indicates how well a document matches the query. In database querying a similar approach is advocated using fuzzy queries [1,3]. Thus, in both cases we have a list of ranked objects (documents/records). A simplest solution is to output the whole ranked list and leave to a user the decision on how many of them are relevant. However, in some cases the decision has to be made automatically by a querying system.

In the paper we propose thresholding strategies based on a linguistic quantifier in the sense of Zadeh. They are both easily interpretable and flexible. In the next section we describe the text categorization problem and the thresholding strategies known in the literature. We point out also the links with similar solutions adopted in database querying. Section 3 serves as a brief resume of the Zadeh's calculus of linguistically quantified propositions. Section 4 discusses proposed thresholding strategies based on this calculus.

2 Text Categorization Problem and Thresholding Strategies

Our understanding of the text categorization problem and a possibility of approaching it using Zadeh's calculus of linguistically quantified propositions is presented in a detail in our previous work [10,11]. Here, we include only brief description of the problem and notation used for the

purposes of further discussion of thresholding strategies.

Let us assume the following notation:

$D = \{d_i\}_{i=1, N}$ - a set of text documents

$C = \{c_i\}_{i=1, S}$ - a set of categories

$f: D \times C \rightarrow [0, 1]$ - assignment of categories to documents by a classifier

$g: D \times C \rightarrow \{0, 1\}$ - final crisp assignment of categories to documents

Thus, we assume we have a classifier (constructed using a training data set) represented by function f that assigns for a given document and a each category a score from the interval $[0,1]$ (for an example of proposed classifier of this type cf. [10,11]). We aim at supplementing such a classifier with a thresholding strategy yielding a crisp assignment of categories represented by function g . Thus, the final result of categorization will be a set of categories characterized by function g . Obviously, each element of this set may be accompanied with a value of function f corresponding to it.

We consider here multiclass multilabel categorization problem, i.e., cardinality of C is higher than 2 and function f may assign non-zero value to a number of pairs (d, c) where d is fixed. Still another dimension making possible to distinguish two subclasses of the categorization problem is the following. We will refer to *on-line categorization* if one document has to be categorized at a time and to *batch categorization* if a set of them is categorized at once. This apparently technical distinction is important for thresholding strategies as discussed below.

Obviously, the higher the score assigned to a category c for a document d , i.e., $f(d,c)$, the more reasonable is to associate the document with the category. However, this score cannot be treated as an absolute measure of the “belongingness” of this document to the category, i.e., there is no absolute optimal threshold on the score values that may be used to decide how to construct function g having function f . In order to solve this problem, the following thresholding strategies are considered in the literature [7]:

- rank-based thresholding (RCut),
- proportion based assignment (PCut),
- score-based local optimization (SCut).

The first strategy consists in choosing r top (i.e., with highest scores) categories for each document. Parameter r may be set by the user or automatically tuned (learned).

The next strategy works for batch categorization and assigns to each category such a number of documents so as to preserve a proportion of the cardinalities of particular categories in the training set.

The last method assigns a document to a category only if a matching score of this category and document is higher than a certain threshold. Thresholds are tuned separately for each category.

In a recent paper Yang [7] proposes two new strategies RTCut and SCutFBR being enhancements of the first and third from among mentioned above, respectively.

Some elements of syntax of fuzzy querying languages proposed in literature (cf. [1,3]) correspond to the SCut thresholding strategy. Both languages are based on SQL and assume that a matching degree of a query against a record is a number from the interval $[0,1]$. Moreover, they propose to extend the SELECT clause of the SELECT instruction with a phrase indicating a required minimal threshold for matching degree of record sought. On the other hand, the RCut strategy bears some resemblance to the classical SQL SELECT instruction requiring top n records. However, the ranking of records is random as all selected records may be treated as being assigned score 1.

In our approach we analyze the question of thresholding strategy from the aggregation operator and flexible constraint perspective. For example, the SCut strategy may be interpreted as selecting a set of categories that all fulfill a condition (their score is higher than a predefined threshold). Taking into account uncertainty/imprecision connected with the scores yielded by a classifier it may be worthwhile to replace such a crisp constraint replacing the aggregation operator exemplified here by the universal quantifier ‘all’ with a more flexible aggregation scheme. In what follows we investigate the possibility to use linguistic quantifier in the sense of Zadeh. That makes it possible to require that, e.g., ‘*most* of the selected categories fulfill a condition’. In order to obtain an operational and reasonable thresholding strategy this has to be supplemented by another condition. We will discuss

that in Section 4, first briefly reminding basics of Zadeh's approach to linguistically quantified propositions.

3 Linguistic quantifiers – Zadeh's approach

The Zadeh's calculus of linguistically quantified propositions [9] offers a formalization of linguistic quantifiers exemplified by "most", "many" "almost all" etc. These may be treated as flexible schemes of aggregation of pieces of information. Basic classical aggregation operators related to logical connectives (AND, OR) and quantifiers (for all, there exists) are often too strict, notably in the context where information to be aggregated is uncertain/imprecise. In many practical situations a human being would express a constraint by stating that, e.g., "Most of the conditions ... should be fulfilled". As it often happens that all/some conditions quantified are of a gradual type, both the conditions and quantifier are best modeled within the framework of fuzzy logic.

Zadeh [9] introduced two types of linguistically quantified propositions:

$$QX's \text{ are } P's \quad (\text{type I}) \quad (1)$$

$$QB's \text{ are } P's \quad (\text{type II}) \quad (2)$$

where Q is a linguistic quantifier, and P and B are fuzzy sets in the universe X . Fuzzy linguistic quantifiers are represented by fuzzy sets defined in an appropriate universe. The *proportional* linguistic quantifiers such as "most", "almost all", etc. are represented by fuzzy subsets, Q , on the interval $[0,1]$:

$$m_Q: [0, 1] \rightarrow [0, 1] \quad (3)$$

Zadeh proposed an interpretation for the proportional linguistic quantifiers such that the truth degree T of proposition (1) is computed using the following formula:

$$T = m_Q\left(\frac{\text{card}(P)}{\text{card}(X)}\right) = m_Q\left(\frac{\sum_i m_P(x_i)}{n}\right) \quad (4)$$

where m_Q is the membership function of quantifier Q and n is the cardinality of the universe X . For propositions of type (2) we have:

$$T = m_Q\left(\frac{\text{card}(P \cap B)}{\text{card}(B)}\right) = m_Q\left(\frac{\sum_i (m_P(x_i) \wedge m_B(x_i))}{\sum_i m_B(x_i)}\right) \quad (5)$$

Thus, the truth of a proposition of type (1) is proportional to the fraction of elements of the universe X that belong to its subset G . An exact form of this relationship is determined by the membership function of Q which may be (for the quantifier "most") of the following, piece-wise linear, form:

$$m_Q(y) = \begin{cases} 1 & \text{for } y \geq 0.8 \\ 2y - 0.6 & \text{for } 0.3 < y < 0.8 \\ 0 & \text{for } y \leq 0.3 \end{cases}$$

The truth of a proposition of type (2) is proportional to the fraction of elements of a (fuzzy) set $B \subseteq X$ that at the same time belong to $P \subseteq X$.

4 Linguistically quantified thresholding strategies

In our previous work [10,11] we proposed a following thresholding strategy. It is trying to exploit the information of co-occurrence of categories labeling training documents. More specifically, for each category we take into account the number of categories with each it is usually assigned to documents; we call them sibling categories. The underlying idea may be expressed as follows:

"Select such a threshold r (rank) that *most of the important categories had a number of sibling categories similar to r in the training data set*"

Thus, for each $r \in \{1, \dots, R\}$ we compute the truth degree of the italicized clause above (R is a parameter) and select such an r for which we get maximum value. This is formalized using Zadeh's calculus of linguistically quantified propositions as:

$$QB's \text{ are } P's \quad (6)$$

where X , the universe considered, is a crisp subset of C of 10 (more generally, R) categories with the highest scores, B is a fuzzy set of important categories for a given document d , i.e., $\mu_B(c_i) = f(d, c_i)$. P is a fuzzy set of categories, that, on average, had in the training set the number of sibling categories similar to r for which truth value of (6) is calculated. This similarity is modeled by a similarity relation which is another parameter of the method.

A new approach proposed here may be expressed as follows:

"Select such a threshold r (rank) that *most of the important categories are selected and most of the selected categories are important*"

Obviously, "most" may be replaced here with another linguistic quantifier. Selection of a suitable quantifier may be a part of the tuning of this thresholding strategy. Although the strategy looks as a variant of RCut it is worth noting its essential difference. It suggests certain rank threshold (as RCut does), but this is done for each document d separately, depending on the whole vector of values $f(d, c_i)$ for all categories.

The proposed strategy may be formalized as a conjunction of two linguistic propositions:

$$QB's \text{ are } P's \quad (7)$$

and

$$QP's \text{ are } B's \quad (8)$$

Operationally, as previously, for each $r \in \{1, \dots, R\}$ we compute the minimum of the truth degrees of both linguistic propositions (7) and (8) and select r maximizing this minimum. As previously, B is a fuzzy set of important categories for a given document d , i.e., $\mu_B(c_i) = f(d, c_i)$. P is a crisp set of categories to be selected, i.e., set of r categories with highest scores.

It may be easily observed that truth value of (7) and (8) for increasing r are non-decreasing and non-increasing, respectively. Thus, the selection of an optimal value of r , maximizing the minimum of (7) and (8) is simple.

Example.

Let $R = 5$ and for certain d

$$f(d, c_1)=1.0, \quad f(d, c_2)=0.8, \quad f(d, c_3)=0.7, \quad f(d, c_4)=0.5, \\ f(d, c_5)=0.0$$

Then, assuming $m_b(x)=x$, we get

r	truth (7)	truth (8)	min
1	0.33	1	0.33
2	0.6	0.9	0.6
3	0.83	0.83	0.83
4	1.0	0.75	0.75
5	1.0	0.6	0.6

Thus, $r = 3$ will be selected.

An interesting feature of the proposed strategy, requiring further study, is a possibility to interpret maximized minimum of (7) and (8) as a measure of confidence in given categorization decision. In the above example we get 0.83 what may be read as a fairly high confidence.

For the study of above possible interpretation of our strategy the following observation may be worthwhile. Namely, the result of the maximization of minimum (7) and (8) may be treated as the Sugeno integral of a function $h: \{1, \dots, R\} \rightarrow [0, 1]$ with respect to a fuzzy measure \mathbf{s} , where

$$h(r) = \text{truth } QA's \text{ are } B's = \\ = m_Q \left(\frac{\sum_{i=1}^r m_B(x_i)}{r} \right)$$

where A is a set of indexes $\{1, \dots, r\}$ and, as previously, $\mu_B(c_i) = f(d, c_i)$. The fuzzy measure \mathbf{s} over the space of all subsets of $\{1, \dots, R\}$ is defined as

$$\mathbf{s}(A) = QB's \text{ are } A's = \\ = m_Q \left(\frac{\sum_{i=1}^R c_A(x_i) \wedge m_B(x_i)}{\sum_{i=1}^R m_B(x_i)} \right)$$

and c_A is a characteristic function of the set A . It may be easily checked that \mathbf{s} verifies axioms of the fuzzy measure.

Both function h and fuzzy measure \mathbf{s} are determined for each document separately and are induced by scores yielded by the classifier used (represented by function f). Function h computes for each index r an average score of r top ranked categories. Fuzzy measure \mathbf{s} for given subset of indexes ("subset of categories") corresponds to the truth of the proposition that "most of highly scored categories belong to that set". This perspective on our new proposed thresholding strategy may shed light on its properties. That will be subject of our further study. Moreover, we plan to carry extensive computational experiments. In our previous work we reported some preliminary results for the experiments with the strategy mentioned at the beginning of this section. Now, we plan to run similar experiments for the new strategy, in a way suggested in [7].

References

- [1] P. Bosc, O. Pivert (2000) SQLf query functionality on top of a regular relational database management system. In O. Pons, M.A. Vila, J. Kacprzyk (Eds.) *Knowledge Management in Fuzzy Databases*. Physica-Verlag, 171-190.
- [2] J. Kacprzyk, G. Pasi, P. Vojtáš, S. ategor (2000). Fuzzy querying: issues and perspectives. *Kybernetika* 6(36), 605-616.
- [3] J. Kacprzyk, S. Zadrozny (2001). Computing with words in intelligent database querying: standalone and Internet-based applications, *Information Sciences* 134, 71-109.
- [4] D.H. Kraft, G. Bordogna, G. Pas (1999). Fuzzy set techniques in information retrieval. In J.C. Bezdek, D. Didier, H. Prade (Eds.) *Fuzzy Sets in Approximate Reasoning and Information Systems*, vol. 3, The Handbook of Fuzzy Sets Series, Kluwer Academic Publishers, Norwell.
- [5] F. Sebastiani (1999) A tutorial on automated text ategorization In A. Amandi, A. Zunino (Eds.) *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence*, Buenos Aires, AR, 7-35.
- [6] Y. Yang (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval* 1(1/2) 67-88.
- [7] Y. Yang (2001). A study on thresholding strategies for text categorization. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*. 137-145.
- [8] Y. Yang, X. Liu (1999). A re-examination of text categorization methods. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 42-49.
- [9] L.A. Zadeh (1983). A computational approach to fuzzy quantifiers in natural languages, *Computers and Maths with Appls.* 9, 149-184.
- [10] S. Zadrozny, J. Kacprzyk (2003) On the application of linguistic quantifiers for text categorization. In *Proceedings of International Conference on Fuzzy Information Processing*, volume 1, 435-440, Beijing.
- [11] S. Zadrozny, J. Kacprzyk (2003) Computing with words for text processing: an approach to the text categorization. *Information Sciences*, to appear.