DEALING WITH TEXTUAL DOCUMENTS IN COPRPORATE APPLICATIONS: TEXT CATEGORIZATION USING FUZZY LINGUISTIC SUMMARIES

SŁAWOMIR ZADROŻNY
JANUSZ KACPRZYK
Instytut Badań Systemowych, Polska Akademia Nauk

Summary

Text documents are the most widely used type of data in corporate practice. Among many aspects of dealing with, and processing of text documents, text categorization is often an initial phase. Text categorization is meant here as an automatic assignment of a text document, characterized by keywords, into some categories. We use a linguistic quantifier guided aggregation, via linguistic summaries, to obtain a concise description of documents, and show its use in text categorization.

Keywords: text categorization, fuzzy linguistic summaries

1. Introduction

Text documents play more and more important a role in corporate practice because an overwhelming majority of data used in organizations and companies is of a textual form. This importance of tech documents is amplified by the fact that a huge amount of text information produced and available in the Internet that is more and more employed as a source of crucial source of data in practical applications.

The above mentioned importance of textual documents and a glut of textual information implies a need to develop tools and techniques that could retrieve information from textual documents that could be useful for various tasks exemplified by data mining and knowledge discovery, decision support, etc.

An initial stage in the utilization of textual documents is their categorization, i.e. their assignment to appropriate categories as, for instance, business, finances, agriculture, industry, etc.

Text categorization may be meant more generally than a pure classification problem. For example, a concise description of a collection of text documents may be valuable. Though at a semantic level this is related to information extraction and text summarization, a a concise description may be useful even when referring to a lower level characteristics of documents (e.g., in terms of keywords, their frequencies of occurrence, etc.). In respect to a concise description of a set of documents, a relatively new and promising technique is the *linguistic summarization*. This paper presents an application of this new technique for the purposes of information retrieval, notably text categorization.

2. Linguistic summarization of data

The purpose of linguistic summarization is to provide means for an intuitive, human-consistent description of a group of objects. A *linguistic summary* may be exemplified by:

"*Many* orders have a *low* commission" (1)

"Most of the young employees have high salary" (2)

and was introduced by Yager (cf., e.g., [22]) and further developed by many authors including Kacprzyk and Yager [9], Kacprzyk, Yager and Zadrożny [10], Kacprzyk and Zadrożny [12].

Formally, such linguistic summaries may be conveniently expressed using Zadeh's [25] *linguistically quantified propositions* as, respectively:

$$Qy\text{'s are } S \tag{3}$$

$$QRy\text{'s are } S \tag{4}$$

with: $Y=\{y_1,\ldots,y_n\}$ - a set of objects to be summarized, e.g. the set of workers. $A=\{A_1,\ldots, A_m\}$ – a set of attributes characterizing y's from Y, e.g. salary. $A_j(y_i)$ is a value of A_j for y_i.

A linguistic summary of a data set Y consists of: (a) a quantity in agreement Q, i.e., a linguistic quantifier (e.g. "*many*"), (b) a summarizer S, i.e., an attribute A_j together with a linguistic term defined on its domain (e.g. "*low commission*" for attribute "*commission*"), (c) optionally, a qualifier R, i.e., another attribute A_k together with a linguistic term defined on its domain determining a (fuzzy subset) of Y (e.g. "*young*" for attribute "*age*"), (d) truth (validity) T of the summary, i.e., a number from the interval [0, 1] assessing truth (validity) of the summary (e.g. 0.7), and may be, therefore, represented as a quadruple (Q, S, R, T).

Using Zadeh's [25] fuzzy-logic-based calculus of linguistically quantified propositions, a (proportional, nondecreasing) linguistic quantifier Q is a fuzzy set in the interval [0,1] as, e.g.

$$\mu_Q(x)=\begin{cases}1 & \text{for } x\geq 0.8\\ 2x-0.6 & \text{for } 0.3<x<0.8\\ 0 & \text{for } x\leq 0.3\end{cases} \tag{5}$$

Then, the truth (validity) of (3) and (4) are calculated, respectively, as

$$\text{truth}(Qy\text{'s are } S)=\mu_Q[\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mu_S(y_i)] \tag{6}$$

$$\text{truth}(QRy\text{'s are } S)=\mu_Q[\textstyle\sum_{i=1}^{n}(\mu_R(y_i)\wedge\mu_S(y_i))/\sum_{i=1}^{n}\mu_R(y_i)] \tag{7}$$

Both, summarizer (S) and qualifier (R) are assumed above in a rather simplified, atomic form referring to just one attribute. They can be extended to cover more sophisticated summaries involving some confluence of various attribute values as, e.g, "*young* and *well paid*".

One can also use other methods of the linguistic quantification as the OWA operators (cf. Yager and Kacprzyk and Dubois et al.'s OWmin, and even generalized quantifiers, cf. [6] or [7]).

Summaries may be formulated "manually" and, possibly, verified autdomatically on the set of objects. In data mining, more attractive is an automatic mining of linguistic summaries. Here, linguistic summaries are meant as succinct descriptions of categories of documents in terms of keywords (terms, tokens). The system produces automatically a set of linguistic summaries for a given set of test documents belonging to a category (categories), but they may be modified/extended by a human operator. We discuss an application based on the vector space model with documents represented as vectors of keywords (their weights in documents).

However, even more attractive from the point of view of manual manipulations of description of categories is a representation referring to the concepts dealt with in a document. In such a case, linguistic summaries may take into account hierarchies of concepts.

Kacprzyk and Zadrożny [11] review a number of approaches for the derivation of linguistic summaries with the use of efficient algorithms for association rules mining which seems to be promising. It makes possible to derive linguistic summaries in a slightly simplified form (basically, both qualifiers and summarizers have to be conjunctions of atomic conditions), however it offers efficient algorithms. In this approach an interpretation of the following generalized and fuzzified forms of association rules are employed, respectively:

$$A_1 \text{ IS } f_1 \wedge ... \wedge A_n \text{ IS } f_n \rightarrow A_{n+1} \text{ IS } f_{n+1} \wedge ... \wedge A_{n+m} \text{ IS } f_{n+m} \quad (8)$$

$$(Q, A_{n+1} \text{ IS } f_{n+1} \wedge ... \wedge A_{n+m} \text{ IS } f_{n+m}, A_1 \text{ IS } f_1 \wedge ... \wedge A_n \text{ IS } f_n, T) \quad (9)$$

with f_i's denoting linguistic terms in the domain of A_i, while Q and T are determined by a so-called *confidence measure* of the association rule (8) – cf. [11] and references therein.

The number of derived linguistic summaries may be huge. We adopt the following pruning scheme for summaries of type (6): a summary (*Q1*, *S1*, *R1*, *T1*) is pruned if there exists another summary (*Q2*, *S2*, *R2*, *T2*) satisfying: *R1* subsumes *R2*, *S2* subsumes *S1*, $Q1 \subseteq Q2$. This leads to a substantial, lossless reduction of the number of rules.

3. Using linguistic summaries in text categorization

Text categorization is a special case of classification. We denote: $D = \{d_i\}_{i=1,N}$- a set of text documents, $C = \{c_i\}_{i=1,S}$ - a set of categories, Ξ: $D \times C \rightarrow \{0,1\}$- assignment of categories to documents, $T = \{t_j\}_{j=1,M}$- a set of terms. Additionally, a set of training documents is considered, i.e., such a set $D_1 \subset D$ that $\Xi(d,c)$ is known for $d \in D_1$ and any $c \in C$

The documents are usually represented by a function: F: $D \times T \rightarrow [0, 1]$, i.e. a document is represented as a vector: $d_i \rightarrow [w_1,...,w_M]$, $w_j = F(d_i,t_j)$, $d_i \in [0,1]^M$ where each dimension corresponds to a term and the value of w_j (weight) determines to what extent a term $t_j \in T$ is important for the description of the document. A popular function F is a *tf×idf* function $F(d_i,t_j)=f_{ij} * \log(N/n_j)$ where f_{ij} is the frequency of a term t_j in a document d_i and n_j is a number of documents containing term t_j. We assume a normalized *tf×idf*. A richer structure of documents may be taken into account as discussed in Section 4.

Often, some additional assumptions are made about C and/or Ξ as, e.g., C contains just two categories and exactly one category is assigned to a document. We adopt here the most general case of multiclass multilabel categorization without these restrictions on C or Ξ.

Many classifier algorithms may be applied, including rule-based systems, decision trees, neural networks, etc, cf., e.g., [17]. Since we need to aggregate partial results obtained when particular terms are taken into account separately, one can apply linguistic quantification. In [27, 29] we used the classic Rocchio algorithm with learning consisting in computing a centroid vector for each category of documents. Then, in classification, a document is classified to a category whose centroid is most similar to this document. As the categories (their centroids) represent many documents, one should not expect a match between a centroid and a document along all dimensions. More reasonable is to formulate a requirement that along *most* of the dimensions there is a match. This may be formalized using the following linguistically quantified proposition: "*A document*

belongs to a category if most of the important terms present in the document are also present in the centroid of the category".

On the other hand, usually the classifiers compute for each document and category a matching degree that yields an ordered list of categories for each document. If just one category is to be assigned, it is natural to choose the one with the highest rank. In multilabel categorization the classifier has to decide how many of top ranked categories should be assigned to a document. This is referred to as a thresholding strategy problem (cf., e.g., [27]). We can: choose a fixed number of top ranked categories for each document; assign such a number of documents to each category so as to preserve a proportion of the cardinalities of particular categories in the training set; or assign a category only if its matching score is higher than a fixed threshold. We proposed some approaches for thresholding strategy [27] based on linguistic quantification. One is to choose such a threshold r that „*most of the important categories had a number of sibling categories similar to r in the training data set*". By a sibling category for a category c_i we mean a category that is assigned to the same document as category c_i. Another approach may be: "*Select such a threshold r (rank) that most of the important categories are selected and most of the selected categories are important*". A rank threshold is selected for each document d separately.

4. Using linguistic summaries for text categorization

Linguistic summaries were originally meant for databases that usually feature a strictly determined structure with clearly identified attributes and their domains. Thus, summaries are well-defined knowing the schema with additional metadata as a dictionary of relevant linguistic terms (cf., e.g., [9, 10]). Text documents usually lack a strict structure and are much less suitable for standard mining techniques, including linguistic summaries. However, as we focus here on the text documents available in the Internet, the situation is more promising because most of such documents reveal some structure, notably since HTML is employed which secures a certain degree of structure of compliant documents (especially XHTML). More and more documents available and exchanged via Internet follow XML specifications which support quite a rich and well defined structure. This makes possible to distinguish different parts of documents that may be important for text categorization. Thus, starting with the *vector space model* and assuming a structure of documents and basic metadata (like name of the file, date of creation, etc.), one has a rich description of text documents.

We assume the model of Internet-based documents by Bordogna and Pasi (cf. e.g., [3]). A document is divided into sections (parts). For typical HTML documents at least TITLE and BODY sections are usually present. Let $P=\{p_k\}_{k\in[1,K]}$ be a set of sections (parts). Then, a document is represented as the vector $d_i=(d_{i11},...,d_{iKN})$, where d_{ikj} denotes the weight of term t_j in section k of document d_i. Thus, d_{ikj}'s are computed by a function F: $D \times P \times T \rightarrow [0, 1]$, and (cf. [3]) it may be proper to use different forms of F for different parts of a document. For example, for a title section, usually containing just a few terms, it may be better to assume the Boolean indexing, i.e. to assign the weight 1.0 to all terms appearing there.

For simplicity, we assume one-level structured documents. Higher-level structured representation (i.e., where a hierarchy of parts is considered) may be more appropriate when a collection of fairly homogeneous semi-structured documents is considered. For higher-level structured documents, Bordogna and Pasi [3] postulate the use of aggregation operators, possibly linguistic quantifiers. That is, when computing F, the formula corresponding to, e.g., *tf×idf*, is used directly only

for triples (d_i,p_k,t_j) such that p_k is a lower level part. For other triples, $F(d_i,p_k,t_j)$ is calculated as an aggregation of $F(d_i,p_l,t_j)$ for all such l's that p_l is nested in (more precisely: is a child of) p_k. Thus, such an approach gives another opportunity for the use of linguistic quantification. In learning algorithms used for classifier construction, linguistic quantifiers (or other aggregation operators) may be tuned (learned) during the training phase.

Thus, we convert a set of original text documents into vectors of term weights from [0,1]. These may be accompanied by information on category (class) belonginess of given document, if available. To reduce the term set we perform typical operations of stopword elimination and stemming. We obtain a counterpart of a set of numerical data typically dealt with using data mining techniques, including linguistic summarization. A text document may be here interpreted as a transaction with terms corresponding to items. Thus, we have a natural setting for the mining of association rules.

We derive linguistic summaries by mining fuzzy association rules. First, we translate crisp data replacing original weights (from [0,1]) with linguistic labels that best match them. We are interested only in terms with high weights, so we use the following dictionary of linguistic labels: "*very important, important, somewhat important*" which are defined as trapezoidal fuzzy numbers on [0.5, 1.0]. The actual shape of their membership functions depends on the distribution of weights in the analyzed set of documents.

Depending on the purpose of linguistic summarization we derive a set of linguistic summaries for a set of documents representing one category or various categories. The former may be useful in case of simple filtering of information, while the latter for a regular categorization of documents. Linguistic summaries derived in the former scenario may also be useful per se – as a description of given set of documents, not necessarily assigned to a specific category. Scenarios in which linguistic summaries are applicable may be summed up as follows.

4.1. Regular text categorization

We also prune the rules which have the same basic form for most of the categories. A "soft" pruning may be done via the reduction of a validity degree of such rules through a formula similar to the *IDF* for terms weighting.

For text categorization the most useful are summaries referring to the category attribute. Moreover, in this context, fuzzy association rules are more obviously applicable – without any further re-interpretation via linguistic summaries. Thus, first of all, rules of the type:

$$A_1 \text{ IS } f_1 \wedge ... \wedge A_n \text{ IS } f_n \rightarrow Category=c \tag{10}$$

are sought.

Derivation of such (crisp) rules for classification is advocated in [13]. They form a set of sufficient conditions for the belongingness of a document to a given category. However, such rules are valid only to some extent – expressed as a confidence measure or a combination of a linguistic quantifier Q and a validity degree T in terms of association rules and linguistic summaries, respectively. Thus, such a rule may be interpreted as setting the lower bound for the belongingness of a document to category c equal to conjunctively combined degrees of truth of the rule antecedent and rule's validity:

$$\mu_c(d) \geq \text{truth}(A_1 \text{ IS } f_1 \wedge ... \wedge A_n \text{ IS } f_n) * T \tag{11}$$

where truth(·) denotes the (degree of) truth of the antecedent of the rule for document d (i.e., how well weights of particular terms in d match labels A_i), and T is validity of the rule. The (11) for the lower bound may be derived by interpreting an association rule $p \rightarrow q$ in the spirit of conditional probability, i.e. truth($p \rightarrow q$) = truth($p \wedge q$)/truth(p), where truth(s) refers to the cardinality of objects verifying s, that is truth($p \rightarrow q$) directly corresponds to the way confidence measure for an association rule is computed. Then, as truth($p \wedge q$) ≤ min (truth(p), truth(q)) (which is obvious in probability theory and also as a general constraint valid for any t-norm $\wedge$) we immediately obtain (11). The formula (11) may also be obtained via the interpretation of an association rule as a fuzzy implication in the sense of Gougen, i.e.,

$$\text{truth}(p \rightarrow q) = \begin{cases} \text{for } p \le q \\ = \text{truth}(q)/\text{truth}(p) \\ \text{elsewhere} \end{cases} \quad (12)$$

Also rules corresponding to the necessary conditions of belongingness, i.e.:

$$Category=c \wedge A_1 \text{ IS } f_1 \wedge ... \wedge A_n \text{ IS } f_n \rightarrow A_{n+1} \text{ IS } f_{n+1} \wedge ... \wedge A_{n+m} \text{ IS } f_{n+m} \quad (13)$$

are worth considering. Obviously, they are also approximate and may be interpreted as setting the upper bound for belongingness of a document to category c:

$$\mu_c(d) \le \text{truth}(A_{n+1} \text{ IS } f_{n+1} \wedge ... \wedge A_{n+m} \text{ IS } f_{n+m}) / T \quad (14)$$

provided that

$$\text{truth}(A_1 \text{ IS } f_1 \wedge ... \wedge A_n \text{ IS } f_n) \ge \text{truth}(A_{n+1} \text{ IS } f_{n+1} \wedge ... \wedge A_{n+m} \text{ IS } f_{n+m}) / T \quad (15)$$

assuming $T \ge 0$, which is trivially verified by association rules usually derived using quite a high confidence measure requirement. If in (13) only the part $Category=c$ appears, then condition (15) should be assumed trivially verified and (14) applies.

Thus, we disjunctively combine the lower bounds produced by rules (10) and conjunctively the upper bounds resulting from relevant rules of type (13). In case of inconsistent lower and upper bounds no conclusion may be drawn about belongingness of a document to a category. In order to avoid such cases, instead of a conjunctive/disjunctive aggregation of bounds some softer, linguistic quantifier guided, schemes of aggregation may be adopted.

4.2. Filtering/routing of documents

One important example of application of the tools and techniques proposed may be to use the linguistic summaries derived (without a category attribute) both to filter out relevant documents and to interactively modify his/her profile. The former is through checking for a new document if it satisfies the linguistic summaries derived. The rules are meant more to *describe* a given category of documents – a primary task linguistic summaries were conceived for – than to *discriminate* among various categories. The latter is limited by operating on the level of terms (keywords). A further development towards linguistic summaries based on concepts and possibly supported by a thesaurus (or ontology) of relevant concepts will here be relevant.

4.3. Comparison of sets of documents

We consider now the problem of how to compare two collections of documents if they are related to the same topic (category, class). The comparison between the summaries derived is possi-

ble but, clearly, more difficult. First, identical summaries are identified. Further, also some relations between summaries (e.g., subsumption) may be exploited.

4.4. Remarks on related works

The use of association rules for classification, a technique used to derive linguistic summaries, was proposed and implemented by Liu et al. [18]. Antonie and Zaïane [1] adopted that approach to text categorization. Hu et al. [8] proposed the use of fuzzy association rules for classification, and Feldman and Hirsh [5] proposed a system (FACT).

5. Concluding remarks

We consider text categorization meant as an automatic assignment of a text document, characterized by keywords, into some categories. We employ a linguistic quantifier guided aggregation, in the form linguistic summaries, to obtain a concise description of documents, and show its use in text categorization. We mentioned that the tools and techniques proposed are relevant in corporate practice in which an overwhelming majority of data items is of a textual character.

Bibliography

1. M.-L. Antonie, O. R. Zaïane (2002). Text Document Categorization by Term Association, in Proc. of the IEEE 2002 International Conference on Data Mining (ICDM'2002), pp 19-26, Maebashi City, Japan, December 9 - 12, 2002
2. G. Bordogna and G. Pasi (2000). Flexible representation and querying of heterogeneous structured documents. Kybernetika, 36(6): 617-633.
3. G. Bordogna and G. Pasi (2003). Flexible representation and retrieval of Web documents. In P.S. Szczepaniak, J. Segovia, J. Kacprzyk and L.A. Zadeh, eds. Intelligent Exploration of the Web, pp. 38-53, Springer-Varlag.
4. D. Dubois, H. Fargier and H. Prade (1997). Beyond min aggregation in multicriteria decision: (ordered) weighted min, discri-min,leximin. In [17], pp. 181 - 192.
5. R. Feldman and H. Hirsh (1996). Mining associations in text in the presence of background knowledge. In Proc. of the Second International Conference on Knowledge Discovery from Databases.
6. I. Glöckner (2003). Fuzzy quantifiers, multiple variable binding, and branching quantification. In T. Bilgic, B. De Baets and O. Kaynak, editors Fuzzy Sets and Systems - IFSA 2003. LNAI 2715, pp. 135 - 142, Springer-Verlag, Berlin and Heidelberg, 2003.
7. P. Hájek, M. Holeňa (2003). Formal logics of discovery and hypothesis formation by machine.Theoretical Computer Science, 292, 345 - 357, 2003.
8. Y.-Ch. Hu, R.-Sh. Chen and G.-H. Tzeng. (2002). Mining fuzzy association rules for classification problems. Computers & Industrial Engineering, 43:735–750.
9. J. Kacprzyk and R.R. Yager (2001). Linguistic summaries of data using fuzzy logic. International Journal of General Systems, 30: 33-154.
10. J. Kacprzyk, R.R. Yager R. and S. Zadrożny (2000). A fuzzy logic based approach to linguistic summaries of databases. International Journal of Applied Mathematics and Computer Science, 10, 4: 813-834.
11. J. Kacprzyk and S. Zadrożny: Linguistic summarization of data sets using association rules. *Proceedings of FUZZ-IEEE 2003 - The IEEE International Conference on Fuzzy Systems*, St. Louis, USA, 2003, IEEE Press, pp. 702-707.

12. J. Kacprzyk J., Zadrożny S.: Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Information Sciences*, 173, 2005, pp. 281-304.
13. J. Kacprzyk J., Zadrożny S.: Towards a synergistic combination of Web-based and data-driven decision-support systems via linguistic data summaries. Springer, Berlin, LECTURE NOTES IN ARTIFICIAL INTELLIGENCE, vol. 3528, 2005, pp. 211-217.
14. J. Kacprzyk J., Zadrożny S.: Towards more powerful information technology via computing with words and perceptions: precisiated natural language, protoforms and linguistic data summaries. In: Nikravesh M., Zadeh L.A., Kacprzyk J. (Eds.): Soft computing for information processing and analysis. Springer Verlag, Berlin Heidelberg New York 2005, pp. 19-33.
15. J. Kacprzyk, Zadrożny S.: Protoforms of linguistic database summaries as a tool for human-consistent data mining. In: Proceedings 14th Annual IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2005), Reno, NV, USA, May 22-25, 2005, IEEE, ss. 591-596.
16. J. Kacprzyk and S. Zadrożny (2005). Fuzzy linguistic data summaries as a human consistent, user adaptable solution to data mining. In B. Gabrys, K. Leiviska and J. Strackeljan, editors, Do smart adaptive systems exist? – Best practice for selection and combination of intelligent methods, Springer, Heidelberg and New York.
17. L.I. Kuncheva (2000). Fuzzy Classifier Design. Physica-Verlag, Heidelberg New York.
18. B. Liu, W. Hsu and Y.M. Ma (1998). Integrating classification and association rule mining. In Proc. of the 4th Int.l Conference on Knowledge Discovery and Data Mining (KDD-98), pp. 80-86, New York, USA.
19. Y. Liu and E.E. Kerre. An overview of fuzzy quantifiers. (I). Interpretations. *Fuzzy Sets and Systems* 95: 1-21, 1998.
20. M. Nikravesh L.A. Zadeh and J. Kacprzyk (Eds.): Soft Computing for Information Processing and Analysis. Springer, Heidelberg and New York, 2005.
21. F. Sebastiani (1999) A tutorial on automated text categorisation In A. Amandi, A. Zunino (Eds.) Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, Buenos Aires, AR, 7-35.
22. R.R. Yager (1996). Database discovery using fuzzy sets. International Journal of Intelligent Systems, 691-712.
23. R.R. Yager and J. Kacprzyk, eds, (1997). The Ordered Weighted Averaging Operators: Theory and Applications. Kluwer, Boston.
24. Y. Yang (1999). An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval, vol. 1, No. 1 / 2, pp. 69-90.
25. L.A. Zadeh (1983). A computational approach to fuzzy quantifiers in natural languages. Computers and Maths with Appls. 9: 149—184.
26. S. Zadrożny and J. Kacprzyk (2003). On the application of linguistic quantifiers for text categorization. In Proc. of International Conference on Fuzzy Information Processing – FIP'03, volume 1, 435-440, Beijing.
27. S. Zadrożny and J. Kacprzyk (2003). Linguistically quantified thresholding strategies for text categorization. In Proc. of the 3rd Int. Conf. in Fuzzy Logic and Technology (EUSFLAT'2003), pages 38-42, Zittau, Germany, September 10-12, 2003.

28. S. Zadrożny and J. Kacprzyk: "An internet-based group decision and consensus reaching support system". In: X. Yu and J. Kacprzyk (Eds.): Applied decision support with soft computing. Springer-Verlag, Heidelberg 2003, pp. 263-275.
29. S. Zadrożny and J. Kacprzyk: On the use of linguistic summaries for text categorization. In: Proceedings of IPMU'2004 – International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (Perugia, Italy), 2004, Vol. 2, pp. 1373 - 1380.
30. S. Zadrożny, K. Ławcewicz and J. Kacprzyk: "Towards an Intelligent Text Categorization for Web Resources: An Implementation". In: B. Bouchon-Meunier, L. Foulloy and R.R. Yager (Eds.): Intelligent Systems for Information Processing: From Representation to Applications. North-Holland, Amsterdam, 2003, ss. 100-111.
31. S. Zadrożny and J. Kacprzyk: "Linguistically quantified thresholding strategies for text categorization". In: Proceedings of EUSFLAT 2003 - Third Conference of the European Society for Fuzzy Logic and Technology, Zittau, Germany, September 10-12, 2003, EUSFLAT and University of Applied Sciences at Zittau/Goerlitz, ss. 38-42.
32. S. Zadrożny S., Kacprzyk J., Gola M.: Towards Human Friendly Data Mining: Linguistic Data Summaries and Their Protoforms. Springer, Berlin, Lecture Notes In Computer Science, No. 3697, 2005, pp. 697-702.
33. S. Zadrożny and J. Kacprzyk: "On the application of linguistic quantifiers for text categorization". In: Proceedings of FIP'03 – International Conference on Fuzzy Information Processing-Theories and Applications, Beijing, China, 2003, Tsinghua University Press/Springer, pp. 435-440.
34. S. Zadrożny, K. Ławcewicz and J. Kacprzyk: TCAT: system automatycznej kategoryzacji internetowych dokumentów tekstowych. In: Piąta Krajowa Konferencja „Inżynieria Wiedzy i Systemy Ekspertowe", Wrocław, 2003, pp. 141-148.

SŁAWOMIR ZADROŻNY
JANUSZ KACPRZYK
Instytut Badań Systemowych, Polska Akademia Nauk
ul. Newelska 6, 01-447 Warszawa, Poland
Tel.:+48228364414, Fax: +48228372772
e-mail: zadrozny, kacprzyk@ibspan.waw.pl