TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
AUTOMATIC CONTROL | AUTOMATYKA
4-AC/2013

SŁAWOMIR ZADROŻNY*, JANUSZ KACPRZYK**, MAREK GAJEWSKI***, MACIEJ WYSOCKI****

A NOVEL TEXT CLASSIFICATION PROBLEM AND ITS SOLUTION

O PEWNYM ZADANIU KLASYFIKACJI DOKUMENTÓW TEKSTOWYCH I JEGO ROZWIĄZANIACH

Abstract

A new text categorization problem is introduced. As in the classical problem, there is a set of documents and a set of categories. However, in addition to being assigned to a specific category, each document belongs to a certain sequence of documents, referred to as a *case*. It is assumed that all documents in the same case belong to the same category. An example may be a set of news articles. Their categories may be sport, politics, entertainment, etc. In each category there exist cases, i.e., sequences of documents describing, for example evolution of some events. The problem considered is how to classify a document to a proper category and a proper case within this category. In the paper we formalize the problem and discuss two approaches to its solution.

Keywords: *text categorization, sequences of documents, sequence mining, hidden Markov models*

Streszczenie

W artykule proponuje się nowe zadanie kategoryzacji dokumentów tekstowych. Podobnie jak w zadaniu klasycznym rozważa się zbiór dokumentów tekstowych i zbiór kategorii. W odróżnieniu od zadania klasycznego, dokumenty są przypisane nie tylko do kategorii, ale również do określonej sekwencji dokumentów w ramach danej kategorii, zwanej *sprawą*. Zakłada się, że wszystkie dokumenty danej sprawy należą do tej samej kategorii. Przykładem może być kolekcja wiadomości prasowych. Mogą one należeć do kategorii takich, jak sport, polityka, rozrywka itp. W ramach każdej kategorii występują sekwencje wiadomości (sprawy) opisujące np. rozwój pewnych zdarzeń. Zadanie polega więc na zaklasyfikowaniu dokumentu do właściwej kategorii i właściwej sprawy w jej ramach. W artykule formalnie definiuje się nowe zadanie kategoryzacji i proponuje się dwa podejścia do jego rozwiązania.

Słowa kluczowe: *kategoryzacja dokumentów tekstowych, sekwencje dokumentów, odkrywanie wzorców sekwencji, ukryte modele Markowa*

* D.Sc. Ph.D. Sławomir Zadrożny, e-mail: slawomir.zadrozny@ibspan.waw.pl, Systems Research Institute, Polish Academy of Sciences, Warsaw; Warsaw School of Information Technology.

** Prof. D.Sc. Ph.D. Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw; Department of Automatic Control and Information Technology, Faculty of Electrical and Computer Engineering, Cracow University of Technology.

*** M.Sc. Marek Gajewski, Doctoral Studies, Systems Research Institute, Polish Academy of Sciences, Warsaw.

**** M.Sc. Maciej Wysocki, Warsaw School of Information Technology.

1. Introduction

Textual documents have always been an important component of broadly meant business processes. Nowadays, more and more of these documents are available in the electronic form. This calls for effective methods of their automatic processing. The relevant research problems are dealt with in the framework of information retrieval (IR). The basic task addressed concerns finding documents satisfying the user's information needs. This requires a proper representation of the documents, information needs (usually in the form of a query) as well as appropriate techniques to match them to each other. Another crucial task is the automatic classification of documents to a set of predefined categories, usually referred to as the text categorization [10, 11]. The categories of interest are often of a topical character, i.e., two documents belong to the same category if they are thematically related. For example, news may be classified in such a way for the purposes of an editorial office of a journal, documents served by a website may be grouped based on their main topics, etc. However, other origins of the categories may be also found in practical applications of the text categorization paradigm. For example, poems may be classified according to their authors, documents received by a company may be grouped according to their type (ads, analyses, reports etc.), or according to their language etc.

Some variants of the categorization task may be distinguished, in a similar way as in the general case of classification. Namely, the binary case where there are just two categories (e.g., of relevant and non-relevant documents with respect to user's preferences, in the *information filtering* task) is of a special interest due to the fact that many well-known classification algorithms are originally meant for the case of two classes. On the other hand, in practical settings the *multiclass* case is much more typical (e.g., item news may be assigned to the politics, economy, sport, etc. category). An important parameter of a text categorization task is the number of categories to be assigned to a document. It may be limited to *at most one*, *exactly one*, or may be unlimited, i.e., a few categories may be assigned to the same document. The latter case is referred to as *multilabel* categorization. Another distinction can be made concerning the mode in which documents are classified, i.e., if they are classified individually, one-by-one (*on-line categorization*), or in groups (*batch categorization*). In the latter mode, the system has more information available while classifying documents but the former mode may be dictated by a practical application at hand.

The text categorization problem, as sketched above, is thus an example of the general classification task. It is most often dealt with in the supervised learning mode. Hence, a training data set is assumed to exist which gathers examples of documents with known class assignment. It is used to construct a classifier which is then used to classify new documents, 'unseen' in the training data set. If the *vector space model* (cf., e.g., [9]) is assumed then documents are represented as numerical vectors making it possible to employ any of the multitude of classifier construction techniques. On an abstract level, these classifier construction techniques may be seen as discovering some regularities characterizing documents belonging to particular categories in the training data set.

In this paper we introduce a new problem of textual documents classification which is an extension of the standard text categorization problem. The problem is inspired by a real life task of the acquisition, maintenance and handling of documents in commercial companies and institutions, with an emphasis on public institutions in Poland. These institutions are

obliged to organize their documents in a precisely specified way. There is a hierarchy (a tree) of topics, so-called JRWA, which comprises some high level topics defined by the appropriate acts of law and low level topics which are adjusted to the specificity of a given institution. Every business process carried out by an institution is assigned to a node of the JRWA tree and all documents related to a given instance of such a process form a *case* (also referred to as a *story*). For example, there may be a JRWA tree node corresponding to the tenders for the equipment purchase. A case is then a particular tender and related documents may comprise a tender announcement, submitted offers, protocols of the tender commission meetings, etc. The documents within a case are chronologically ordered according to the date a document has been created or received.

The classification problem considered in this paper may thus be briefly described as follows. Let us assume a JRWA tree with a number of cases assigned to its nodes. Usually these cases will be at a different stage of development. For example, one tender may have been just announced and its list of related documents consists of only the announcement, while other tenders may be close to their closing. The same applies to the cases gathered in other nodes of the JRWA tree. The problem which we face is a proper classification of a document which has been just received or produced. It has to be classified both to an appropriate JRWA node and to a specific case within this node. The former classification is somehow easier and close to the standard text categorization problem. The latter is much more complex and requires taking into account the relationship between subsequent documents in the chronological order of particular cases. Both classification problems are intermingled as, first of all, assigning a document to a case implies its assignment to the JRWA node to which this case belongs. On the other hand, such a direct classification to a case may be difficult and preceding it by first classifying a document to a JRWA node may be advantageous.

In the following sections we formally state a new text categorization problem and briefly discuss its possible solutions.

2. Formal statement of the new text categorization problem

We assume a collection of documents arranged in ordered sequences, referred to earlier as cases, which are assigned to nodes of a hierarchy of categories. We will adopt the following notation:

- $D = \{d_1, ..., d_n\}$ is a set of documents,
- $C = \{c_1, ..., c_m\}$ is a set of categories of documents, arranged in a hierarchy (tree),
- $\sigma_k = \langle d_{k1}, ..., d_{kl} \rangle$ is a *sequence of documents* (*case*) of documents,
- $\Sigma = \{\sigma_1, ..., \sigma_p\}$ is a set of cases; all documents of a case belong to the same category or, equivalently, each case is assigned to a category; all cases are pairwise disjoint.

Thus, there are two orthogonal classification schemes in place here. We assume that documents in the same category are somehow similar thematically while documents belonging to the same case form some logical sequence. For example, let us consider a set of news articles. The categories here may be *sport*, *politics*, *entertainment,* etc. In each category there exist sequences of documents (cases) describing, for example evolution of some events. For instance, in the *sport* category there may be a case for the Olympic games, another one for football world championships, etc. The documents (articles) in the former

case may report results of consecutive matches/competitions while in the latter subsequent articles, may discuss stages of the preparations for the championship. Of course, these are just examples and a different hierarchy of topics and interpretation of cases may be assumed implying a different arrangement of the articles.

Let us now consider a new document d^*. By definition it belongs to a case in some category. Our goal is to devise a way of constructing a classifier which will suggest the assignment of d^* to a proper case.

In the next sections, we propose two ways to construct such a classifier. Both approaches follow the standard framework of the supervised learning. A collection of documents D, arranged in a number of cases, Σ, is assumed to be available which is split into a training and testing part. The testing part has to be further split as for testing we need:

– a set of cases at various stages of evolvement, forming a structure in context of which the documents have to be classified,
– a set of individual documents which have to be classified with respect to the above mentioned testing structure.

Let us denote the original testing part of the collection of documents as $\Sigma_T = \{\sigma_1, ..., \sigma_p\}$. The testing structure (set of cases), $\Sigma' = \{\sigma'_1, ..., \sigma'_w\}$ is constructed in such a way that $w \leq p$ and if $\sigma'_k \in \Sigma'$, $\sigma'_k = \langle d_{k1}, ..., d_{km} \rangle$ then there have to exist $\sigma_k \in \Sigma_T$, $\sigma_k = \langle d_{k1}, ..., d_{kl} \rangle$, where $m \leq l$. Thus, in general, not all cases of Σ_T appear in the testing structure Σ', and those which do appear are, in general, subsequences of a certain number of initial documents of the original cases from Σ_T. The set of individual documents to be a subject of the test classification comprises all documents omitted in Σ'. It is assumed that the documents are presented to the system in 'chronological' order, i.e., if $\sigma_k = \langle d_{k1}, ..., d_{km}, ..., d_{kl} \rangle$ and $\sigma'_k = \langle d_{k1}, ..., d_{km} \rangle$ then d_{km+i} has to be classified before d_{km+i+1}. Of course, due to the above described construction, for each test document a proper case to which it should be assigned is known.

The problem of classification documents to cases is more difficult than the standard text categorization problem. It is not enough to decide which category a document belongs to. It has to be attached to an existing case, at its end, or should initiate a new case. Thus, the essence of the classifier construction is to learn some rules linking documents in cases of a given category. It should be noted that such a classification is carried out manually at the companies by their employees. Their job is somehow easier as they can understand the meaning of the document and take advantage of the metadata usually accompanying the document (the date, the addressee and sender, some reference in the document to other documents belonging to its case etc.). In particular, documents which are generated by a given company are usually clearly related to some cases. Hence, our formulation of the problem is more general and requires a classifier to make a decision based only on the content of the given document and the knowledge of the characteristic features of cases in particular categories, learned from a training data set. However, such a general formulation still applies to many practical problems. For example, if the legacy of a person has to be organized then the metadata available may be scarce and there are practically no additional clues except the content of the documents.

The novelty of the problem is related to the need to combine two perspectives: a standard text categorization perspective and a case-based classification one. The former problem has been thoroughly studied in the literature and various well-known techniques of supervised learning have been applied to solve it. The latter classification problem has not yet been clearly and explicitly identified in the literature and only some slightly related formulations

and solution have been proposed, cf., e.g., the Topic Detection and Tracking (TDT) problem [2]. Here we are mostly interested in dealing with this latter classification problem but the first formulation will also play an important role in our considerations, and – from a more general perspective – a synergistic combination of these two problems and their solutions is a real challenge to be tackled.

3. A Hidden Markov Model based approach

In what follows, we assume that the documents in question are represented as vectors over an appropriate space. The dimensions may correspond to particular *keywords* (*terms*) from a set T, $t = \{t_1, ..., t_m\}$, as in the classical vector space model (cf. [3]), to the *topics* identified using the Latent Dirichlet Allocation modelling or to any other entities used in various approaches to the modelling of documents within the information retrieval realm.

We want to model the succession of documents in case, specific for particular categories. Intuitively, a subsequent document in a case may be treated as corresponding to a step in the development of a given case. Referring to a previous example of a case meant as a tender for the purchase of equipment, we can distinguish such steps as: preparation of the terms and conditions of the tender; receiving questions of potential providers and answering them; receiving offers; making a choice of the provider; preparing a contract etc. The steps may overlap in time, e.g., some potential providers may submit they offers while others may still pose questions concerning the terms and conditions. Thus, the succession of documents has a probabilistic character and, moreover, the steps cannot be directly identified based on the documents themselves. Hence, for modelling the cases, we propose to employ Hidden Markov Models (HMM), cf., e.g., [7]. We distinguish the following elements:

- *hidden states* $S = \{S_1, S_2, ..., S_L\}$ which may be interpreted in the context of our problem as corresponding to particular stages of a given case type; e.g., various steps in the tender procedure mentioned above; a state in a time moment t will be denoted as q_t,
- *observations* generated by an HMM in subsequent steps, corresponding here to the whole documents (in fact, vectors representing them) d forming a case; thus a multidimensional continuous space of observations is assumed here,
- *a state transition matrix* $A = [a_{ij}]$ defining the probability of going from one state to another, $a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, $1 \le i, j \le L$,
- observation probability distributions b_j defined for each state j in the space of documents D; we will assume that these distributions are mixtures of M normal distributions: $b_j(d) = \sum_{m=1}^{M} w_{jm} \mathrm{N}(d, \mu_{jm}, U_{jm})$, $\sum_{m=1}^{M} w_{jm} = 1$, $w_{jm} \ge 0$, $1 \le j \le L$, where w_{jm} is a weight of the m-th component of the mixture for a state S_j, $\mathrm{N}(d, \mu_{jm}, U_{jm})$ is a normal distribution in the space of document vectors with the mean vector μ_{jm} and the covariance matrix U_{jm}; a vector of all such distributions will be denoted as B; a matrix of the weights of the components of the mixtures for particular states will be denoted as W; set of the matrices of the mean vectors of these components will be denoted Ξ, and a set of the matrices of corresponding covariance matrices will be denoted Ψ,
- an initial probability distribution in the space of states, $\pi = [\pi_1, \pi_2, \ldots, \pi_L]$ where $\pi_j = P(q_1 = S_j)$, $1 \le j \le L$.

A separate training collection of complete cases for each category is assumed to be available. We start by setting a number of states L and then a separate HMM, $\lambda_c = (A_c, B_c, \pi_c, W_c, \Xi_c, \Psi_c)$, for each category c is trained which results in the learning of the probability distributions A_c, B_c and π_c as well as the parameters Ξ_c and Ψ_c of the mixtures and normal distributions mentioned above. The standard EM based algorithm is used to train each HMM [7].

The classification of a new incoming document d^* is carried out as follows. A set of on-going cases forming a current collection of documents arranged in cases for each category is available. For each such case $\sigma = (d_1, d_2, ..., d_p)$ a *matching degree md* of the document d^* against the case σ is computed as the conditional probability that the HMM λ_c will generate the sequence of documents currently forming the case extended with the document d^* added at its end:

$$md(\sigma, d^*) = \mathrm{P}_\sigma(d_1, d_2, ..., d_p, d^* \mid d_1, d_2, ..., d_p, \lambda_c) = \frac{\mathrm{P}(d_1, d_2, ..., d_p, d^* \mid \lambda_c)}{\mathrm{P}(d_1, d_2, ..., d_p \mid \lambda_c)} \tag{1}$$

For each category an 'empty' story is also considered and then (1) takes the following form:

$$\mathrm{P}_\sigma(d^* \mid \lambda_c) = \sum_{j=1}^{L} \pi_j b_j(d^*) \tag{2}$$

The document d^* is assigned to such a case σ^* that:

$$\sigma^* = \arg\max_{\sigma} md(\sigma, \theta^*) \tag{3}$$

There is a number of parameters which have to be set before the above described classification can be carried out. Among them, the number of states of HMMs representing cases in particular categories have to be determined. They may be set experimentally and/or via the analysis of a collection at hand. For instance, the number of states may be related to the average length (number of documents) of completed cases in a given category.

4. A sequence mining based approach

The second approach to solving the new text categorization problem we propose in this paper consists of using the *sequence mining* approach [1, 15] to model logical relations between the documents in a case of within a given category. Let us adopt the following notation [15]. Documents are now treated as sets of keywords, $d_i \subseteq T$, and the task of sequence mining boils down in our context, to finding *sets of keywords* $f_i \subseteq T$ frequently appearing in stories of a given category.

Let $F = \langle f_1, f_2, ..., f_r \rangle$ and $G = \langle g_1, g_2, ..., g_s \rangle$ denote sequences of sets of keywords. The sequence F is said to be a *subsequence* of sequence G, denoted as $F \prec G$, if there exists such an injection $h,\sim h : \{1, 2, ..., r\} \rightarrow \{1, 2, ..., s\}$, such that:

$$\forall_{f_i \in F} \; f_i \subseteq g_{h(i)} \wedge \left((i < j) \Rightarrow \left(h(i) < h(j) \right) \right) \tag{4}$$

It should be noted that a sequence of documents (a case) σ may be viewed to be a sequence of sets of keywords because we assume here that a document is represented by a set of keywords. It is then said that a case σ *contains* a sequence of sets of keywords F if $F \prec \sigma$. The *support of a sequence of sets of keywords F* in a set of cases Σ is defined to be the number of cases containing sequence F, which may be denoted as:

$$supp\left(F, \Sigma\right) = \left\{ \sigma_i \in \Sigma | F \prec \sigma_i \right\} \tag{5}$$

where $\|\cdot\|$ denotes the cardinality of the corresponding set.

A sequence of sets of keywords is said to be *frequent* in a given set of stories Σ if its support is greater than some threshold value *min_supp*:

$$supp\left(F_{cz\square sta}, \Sigma\right) \geq min\_supp \tag{6}$$

There are many algorithms, exemplified by SPADE [15], which make it possible to discover all frequent sequences of sets of keywords for a given set of cases.

Using frequent sequences one may determine *rules* describing dependencies between the occurrence of particular sets of keywords. For example, if:

$$\begin{gathered} supp\left(F, \Sigma\right) = x \\ supp\left(G, \Sigma\right) = y \\ F \prec G \end{gathered} \tag{7}$$

then it is said that the rule:

$$F \Rightarrow G$$

holds with a confidence level *conf* equal:

$$conf\left(F \Rightarrow G\right) = \frac{y}{x} \tag{8}$$

We are interested in particular in so-called *strong rules* $F \Rightarrow G$ such that:

$$conf\left(F \Rightarrow G\right) \geq min\_conf \tag{9}$$

if:

$$F = \langle f_1, f_2, ..., f_r \rangle \quad \text{then} \quad G = \langle f_1, f_2, ..., f_r, g_{r+1} \rangle \tag{10}$$

where *min_conf* denotes some required minimal level of the rule confidence.

Such rules will be denoted in a simpler form as:

$$F \Rightarrow g_{r+1} \tag{11}$$

We will now describe how frequent sequences are used to solve our new text categorization problem. During the training phase a collection of cases for each category is assumed to be available. This time, as mentioned earlier, a document is represented by a set of keywords. For each collection of cases, frequent sequences are discovered and their corresponding strong rules are generated, cf. (11).

To classify a new document d^*, we proceed as follows:

1) for each case σ all *active rules* are considered which match this case, i.e., such rules in which the left hand side sequence of sets of keywords F is a subsequence of the case $\sigma : F \prec \sigma$; for a given case, only the rules generated for the category to which this case belongs are taken into account;
2) among the rules $F \Rightarrow g_{r+1}$ we count those for which the right hand side g_{r+1} is a subset of the document to be classified, i.e., $g_{r+1} \subseteq d^*$ (it should be noted that both the documents and the right hand sides of the rules are sets of keywords); rules having the same right hand sides are counted only once;
3) the document d^* is classified to such a case for which the number of rules counted in step 2 is the highest, and also higher than a certain threshold value *min_count*; if there is more than one such a case, then one of them is randomly selected;
4) if there is no such case for which the number of rules counted as in step 2 is higher than *min_count*, then such a document d^* starts a new story in the category which is selected using a standard text categorization algorithm, e.g., based on the Naïve Bayes approach [6, 13].

In the next section, we present the results of some computational experiments using this algorithm.

5. Computational experiments

In section 2, we have formally introduced a new text categorization problem. There are no standard datasets to test the proposed methods of its solution. Thus, we adapted the ACL Anthology Reference Corpus (ACL ARC) [4] for our purposes. It consists of 10291 scientific papers on computational linguistics. In our preliminary experiments, we employed a subset of this corpus. Each paper comprises a number of explicitly distinguished sections. For our purposes we identified each paper with a sequence of documents σ and its sections are particular documents in such a sequence. This way we obtained 113 sequences, consisting of 11 documents on average.

All documents were represented using the vector space model [3] and the tf x IDF weighting scheme with normalization with respect to the vector length, in particular. Standard document processing techniques were applied, such as stopwords elimination and stemming.

The documents were grouped into categories using the k-means algorithm based cluster analysis. The number of categories was chosen experimentally to be equal 7. Two clusters were ignored due to their small cardinality and finally, a set of 98 document sequences was obtained.

The cSPADE algorithm was used to generate a set of rules (11) for each category. The arulesSequences package for the system R [8] was employed. For sequences mining the representation of each document was limited to the 10 upper-most keywords with respect to their tf x IDF weight in a given document.

The classification algorithm was run four times, each time randomly selecting the test set of documents using the procedure described in section 2. The results of these four runs are briefly presented in Table 1.

Table 1

Results of the computational experiments with the algorithm based on sequences mining

Run No.	Total number of classified documents	Number of correctly classified documents	Microaveraged precision	Macroaveraged precision over the categories
1	252	208	0.8254	0.8531
2	245	204	0.8326	0.8576
3	241	197	0.8174	0.8481
4	252	201	0.7976	0.8101

In virtually all runs, precision of at least 80% was obtained. The results are thus encouraging but the experiments have to be continued with larger data sets as well as with real data sets of cases. Such data are not easy to get but we are in the process of building a collection of documents of one of the public administration institutions in Poland.

6. Conclusions

We have defined a novel and extended text classification related problem that combines issues relating to the acquisition, maintenance and handling of documents in a corporate and institutional setting. We proposed a formal statement of the problem and two approaches to solve it. A pilot implementation of one of the algorithms has been implemented and some preliminary computational experiments hve been carried out [12]. The results obtained are promising but some further experiments are needed to confirm the effectiveness of the proposed method.

This research was partially supported by the National Science Centre (contract No. UMO-2011/01/B/ST6/06908).