

Improving Hungarian Text Categorization Using Domain-Specific Ontology

István Pilászy^a, András Föhrécz^b

^aDepartment of Measurement and Information Systems,
Budapest University of Technology and Economics
e-mail: pila@mit.bme.hu

^bDepartment of Measurement and Information Systems,
Budapest University of Technology and Economics
e-mail: fandrew@mit.bme.hu

Abstract

The aim of Text Categorization is to automatically assign documents to a set of predefined categories. The prevailing approach is making use of a collection of precategorized examples for the induction of a document classifier through learning methods. In this paper we introduce a method which combines state-of-the-art learning techniques with background knowledge. We have used KAON ontology for knowledge representation. We have developed a reasoning method which makes use of the relations in the ontology. Our experiments will show that the method substantially enhances the results of text categorization, it will be clear that a domain specific ontology can improve performance. The proposed method is applicable in the field of spam filtering, document reorganization and classifying news stories and e-mails.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3: Information Search and Retrieval;

Key Words and Phrases: Information Retrieval, Background Knowledge, Ontology, KAON, Text Categorization, Text Classification.

1. Introduction

Nowadays through the sudden growth of the Internet and on-line available documents, the task of organising text data becomes one of the principal problems. A major approach is text categorization (TC), which is used to classify news stories, to filter out spam and to find interesting information on the web. Until the late '80s

the most popular methods based on knowledge engineering, i.e. manually defining a set of rules encoding expert knowledge. In these days the best TC systems use the machine learning approach: the classifier learns rules from examples, and evaluates them on a set of test document.

In this paper we propose a new approach which introduces the use of background knowledge in text categorization. We have used a domain-specific ontology for representing knowledge, and have developed a reasoning mechanism, which associates relevant entities. Our method takes the pre-processed documents, splits them into sentences, filters out words not in the ontology, associates new ones according to the relations in the ontology, and the result is the input for the text classifier.

We have evaluated the method on a Hungarian text corpus, which have resulted in a substantial improvement in performance and generalization, i.e. from a small set of examples the enhanced classifier performs much better than the original from a larger set.

1.1. Steps of text categorization

Text categorization consists of text pre-processing, mapping into vector space, machine learning and testing. Pre-processing usually means stopword filtering for omitting meaningless words, word stemming for reducing the number of distinct words, lowercase conversion etc.

After pre-processing steps the documents will be mapped into a vector space. The vectors together form the term-document matrix. We apply the most commonly used TF•IDF term-weighting method [1]. It is desirable that documents of different length have the same length in the vector space, which is achieved with the so-called document normalization.

The dimensionality may be very high, which is disadvantageous in machine learning, thus dimension reduction techniques are called for. Two possibilities exist, either selecting a subset of the original features, or transforming the features into new ones, that is, computing new features as some functions of the old ones.

After pre-processing and transformations, a machine learning algorithm is used for learning how to classify documents. Support Vector Machines (SVM) have been proven as one of the most powerful learning algorithms [2], and they perform well with high dimensional data, too [3].

For Hungarian texts the TC task is harder because of the lack of publicly available natural language processing technologies, the very few respective publications, and the complexity of the language. The Hungarian language is an agglutinating language like Finnish or Turkish, so it is hard to get a stem of a word.

2. USING BACKGROUND KNOWLEDGE

Background knowledge is a collection of *a priori* knowledge, which helps to solve an information retrieval problem more accurately and efficiently. In the case

of text categorization it is expected to help to find the most relevant features, based on the semantics. We should select such a knowledge representation scheme, which is the most easily usable in this aspect.

When choosing between possible knowledge representations, we should examine the roles they can fill in, and select the most important ones [4]. To cover the domain precisely, we need a *medium for human expression*; to utilize the described knowledge, a *fragmentary theory of intelligent reasoning* is the relevant role. Ontology, the chosen representation, is an easy way of visualizing semantic relations and provides flexible reasoning capabilities. Ontology is a description of the concepts and relationships that exist in some domain [5].

Staab *et al.* propose an ontology-based document clustering method [6]. They compare traditional term-based indexing, the use of a term selection method, and finally the clustering based on the view created from the ontology. However, in the ontology only the sub-class-of relation is introduced, using only taxonomy to represent background knowledge. We would use a more expressive scheme to recognize text semantic more accurately.

For testing and comparing text categorization methods, we have chosen a corpus containing more than 2200 articles about light music, found on <http://music.hu/> Hungarian music portal. The articles are already classified into eight categories.

2.1. Creating an ontology

From the diverse set of ontology languages we have chosen KAON¹, which has enough expressiveness, is flexible and offers a comprehensive tool suite [7]. The tools are open-source, making possible to implement the reasoning machine conveniently.

In absence of reusable Hungarian sources, we examined English sources as well. Unfortunately freely available ones² are not complex enough for our purpose, consisting only of simple taxonomies or enumerations. So we had to start ontology development from scratch. Our starting point was the set of keywords extracted from the corpus by hand, and we expanded the ontology by covering gradually the whole domain.

2.2. Reasoning

The purpose of reasoning is to expand the document index with relevant features, using the relations in the ontology. Starting from the entity identified in the document, we associate with the related entities. Considering the semantic of the ontology we reason to the most appropriate ones only.

The weighted expansion of the document index is done in five distinct steps, using different rules to the various kinds of relations:

Instance-of edges: we associate the parent concepts from the identified instances. For example we generalize from *Metallica* to the *band* concept.

¹KAON Ontology and Semantic Web Infrastructure <http://kaon.semanticweb.org>

²Music Domain Ontology in DAML Ontology Library <http://www.daml.org/ontologies/276>

Concept hierarchy: we should deduce the more general concepts, always reasoning upwards within the hierarchy (e.g. *band* → *performer*). The higher distance means weaker relation, which is realized with a geometric sequence of the distance.

Property hierarchy: the method is the same, as in the case of concept hierarchy

Domain and range edges: a property and the connected concepts represent a close semantic relation, in many cases they correspond to a syntactic structure. The reasoning procedure tries to recognize these and associate the missing parts (e.g. *performer*, *sing* → *song*).

Property-instances: describe relations between instances, help to deduce facts about them (e.g. *Scooter* → *German*). The weight of the appropriate property is taken into account, because relates to the relevance of the property-instance edge.

The order of the steps can be determined by the examination of the expected behaviour, which steps should precede others. Fortunately every influence explained by semantics can be obtained by a proper order of execution. There's no need for complicated recursions, the structure of reasoning remains simple.

3. Experiments

We would like to measure, how the ontology based indexing influences the accuracy of the categorizer. In order to evaluate different aspects of the method, we introduce four categorization engines, adding certain steps to the traditional model gradually. (Fig. 1. shows the methods and a Hungarian example processed by each of them.)

First, we use a traditional **term based indexing** model. As the complexity of Hungarian language requires, word stemming is used, and thereafter stopwords are filtered. After using TF-IDF weighting, the document vectors are normalized.

The **wordlist based indexing** keeps only those words, which are included in the ontology as a stem of an entity. This is similar to stopwords filtering, with only keeping words that are considered relevant.

Synonym based indexing merges words with synonym meaning, by replacing them with the corresponding entity from the ontology. There may be words with two or more related entities, but we don't use word sense disambiguation (WSD) techniques.

Ontology based indexing appends entities to the results of synonym-based method with the discussed ontology-based reasoning.

3.1. Implementation

We stemmed the words with MorphoLogic's Humor³ (High-speed Unification MORphology). Afterwards we used an own list of stopwords to filter out meaningless words. The training document set contained at most 1400 documents and the

³MorphoLogic – Humor http://www.morphologic.hu/en/en_humor.htm

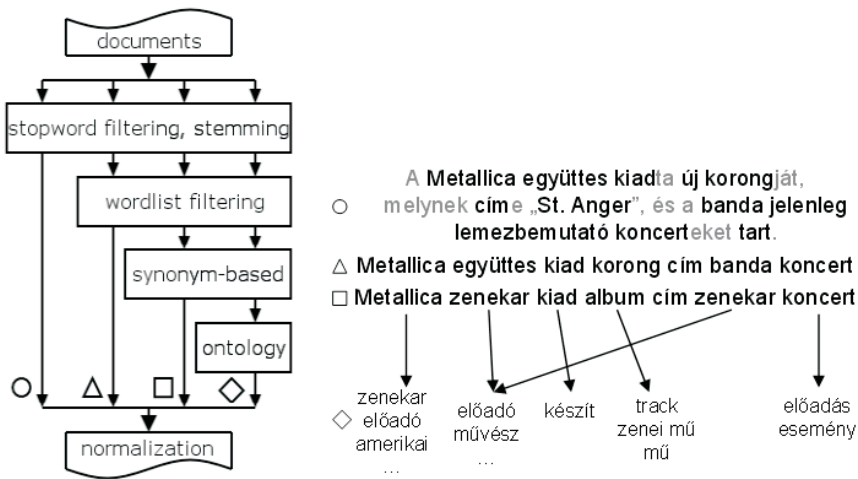


Figure 1: Methods of document indexing

test set 500 documents. We built our ontology with KAON OI-modeler. At the reasoning step, we used the KAON API to access and manipulate the ontology.

As the classifier, we used an SVM implementation called SVM^{light}⁴, which is well suited for large scale and sparse problems. In all experiments we used a linear kernel. A binary classifier was trained for each of the eight classes by using each class at a time as positive examples with the rest of the data as negative examples.

The reasoning algorithm can be applied for texts of different sizes. The articles were originally split into paragraphs, and we split them into sentences. The reasoning was tested on the original articles, on the paragraphs and on the sentences. The best results were achieved by the sentences.

3.2. Results

We have compared our approach with the term-, wordlist- and synonym-based indexing schemes. We have tested these four methods on ten training document sets of different sizes, and evaluated them using the micro and macro averaged F₁ measure (the higher the better), and the optimal dimensionality of the vector space (the lower the better) using the terms with the highest document-frequency (number of documents in which the term occurs).

The F₁ measure is the harmonic mean of precision and recall. For one category, precision is the ratio of truly classified positive examples and positively classified examples. Recall is the ratio of truly classified positive examples and the number of documents in the category. For multiple categories, the precision and recall may

⁴SVM^{light} <http://svmlight.joachims.org/>

be micro- or macro-averaged. In case of micro-averaging we take category sizes into account, as opposed to macro-averaging [8].

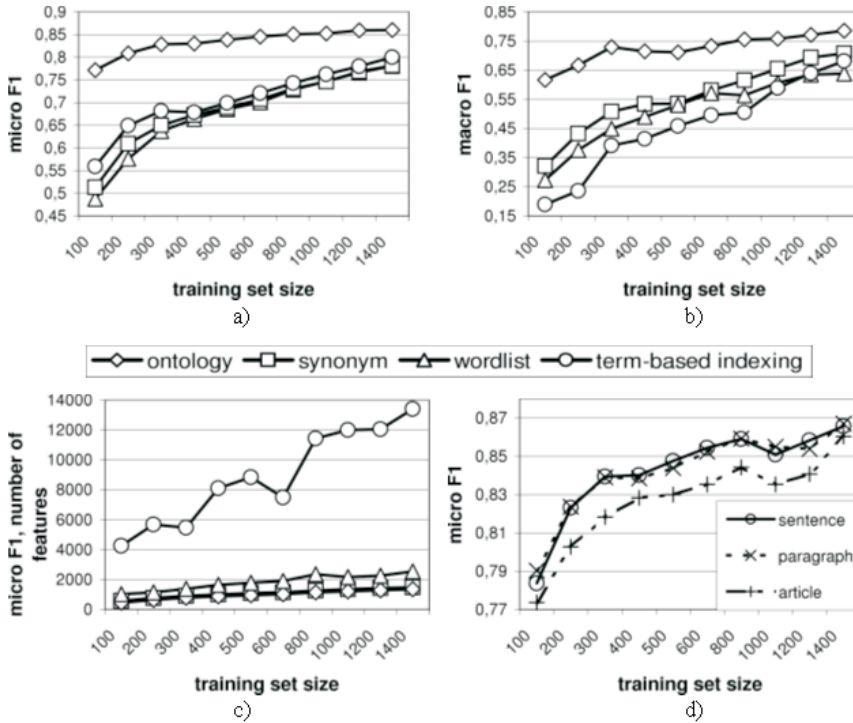


Figure 2: Results

Results are depicted in Fig. 2. The horizontal axis denotes the number of documents the training set contains. The micro and macro F_1 are represented by the vertical axis of the chart. There is a remarkable improvement as a result of our ontology-based indexing scheme. Its ability to generalize is excellent: from a small number of training documents it gets very good results and the results hardly improve by increasing the size of the training set. In micro F_1 the other three schemes perform approximately equally. This means, that our ontology contains the important words, and SVM is good in overlearning. In macro F_1 these three methods differ more, because the very small categories are emphasized compared to micro F_1 , and the dimensionality is too high for the SVM to learn the relevant rules. For this reason it is not surprising, that the synonym-based method is better than the other two.

The vertical axis on Fig. 2/c denotes the optimal dimensionality of the vector space. Our ontology-based indexing scheme achieves the best results keeping only the tenth of features compared to the term-based scheme.

We have compared the ontology based reasoning on sentences, paragraphs and on the original articles. Results are depicted on Fig. 2/d. There is no remarkable difference between sentence- and paragraph-based reasoning, but both of them improve the results compared to the reasoning based on the whole articles.

4. Future work

We are going to deal with word sense disambiguation (WSD). There are a lot of publicly available heuristics for this task [9], but we would like to use all the entities and edges of our ontology. Furthermore we want to examine how our method is applicable for bigger, more general ontologies. It should be studied, how the terms with the highest and lowest weights ranked by the linear SVM at the learning step are usable for ontology-building.

5. Discussion

In this paper, we presented a novel approach for text classification. We introduced a method combining state-of-the-art techniques with background knowledge. We used KAON ontology for knowledge representation, and suggested a heuristic reasoning algorithm which makes use of the entities and edges of the ontology.

Our experimental evaluation has shown that the method provides substantially better results of text classification. We compared the term-, wordlist- and synonym-based methods to our approach. From the results it is clear that a domain-specific ontology can improve performance.

References

- [1] A.Aizawa, "An information-theoretic perspective of tf-idf measures", *Information Processing and Management: an International Journal archive*, Vol. 39, Issue 1, pp. 45–65, 2003.
- [2] Fabrizio Sebastiani, "Machine learning in automated text categorization", *ACM Computing Surveys*, vol. 34, no. 1, pp.1–47, 2002.
- [3] Thorsten Joachims, "Text categorization with support vector machines: learning with many relevant features", *Proc. of ECML-98, 10th European Conference on Machine Learning*, pp. 137–142, Springer Verlag, Heidelberg, DE, 1998.
- [4] Randall David, Howard Shrobe, Peter Szolovits, "What is a Knowledge Representation?", *AI Magazine*, vol. 14, no. 1, pp. 17–33, 1993.
- [5] Nicola Guarino, Pierdaniele Giaretta, "Ontologies and Knowledge Bases", *Mars N.J.I. (ed.): Towards Very Large Knowledge Bases*, pp. 25–32, IOS Press, Amsterdam, 1995.
- [6] Steffen Staab, A. Hotho, "Ontology-based Text Document Clustering (Extended Abstract of Invited Talk)", *Proc. of the Conference on Intelligent Information Systems*, Physica/Springer, Zakopane, Poland, 2003. July

- [7] B. Motik, A. Maedche, R. Volz, “A Conceptual Modeling Approach for Semantics-Driven Enterprise Applications”, *Proceedings of the First International Conference on Ontologies, Databases and Application of Semantics (ODBASE-2002)*, Springer, California, USA, 2002.
- [8] David D. Lewis, “Evaluating Text Categorization”, *Proc. of Speech and Natural Language Workshop*, pp.312-318, 1991.
- [9] Nancy Ide, Jean Véronis, “Word Sense Disambiguation: The State of the Art”, *Computational Linguistics*, 24(1), pp. 1–40, 1998.

Postal addresses

István Pilászy

*Department of Measurement
and Information Systems,
Budapest University of Technology
and Economics,
Bükkfa utca 13,
H-1028 Budapest
Hungary*

András Förhécz

*Department of Measurement
and Information Systems,
Budapest University of Technology
and Economics,
Király utca 51. IV/35,
H-1072 Budapest
Hungary*