Social Semantics And Its Evaluation By Means Of Closed Topic Models: An SVM-Classification Approach Using Semantic Feature Replacement By Topic Generalization

Ulli Waltinger, Alexander Mehler and Rüdiger Gleim

Bielefeld University, Goethe University Frankfurt
{ulli_marc.waltinger,alexander.mehler}@uni-bielefeld.de,
gleim@em.uni-frankfurt.de

Abstract. Text categorization is a fundamental part in many NLP applications. In general, the Vector Space Model, the Latent Semantic Analysis and Support Vector Machine implementation have been successfully applied within this area. However, feature extraction is the most challenging task when conducting categorization experiments. Moreover, sensitive feature reduction is needed in order to reduce time and space complexity especially when deal with singular value decomposition or larger sized text collections. In this paper we examine the task of feature reduction by means of closed topic models. We propose a feature replacement technique conducting a topic generalization comprising user generated concepts of a social ontology. Derived feature concepts are then subsequently used to enhance and replace existing features gaining a minimum representation of twenty social concepts. We examine the effect of each step in the classification process using a large corpus of 29,086 texts comprising 30 different categories. In addition, we offer an easy-to-use web interface as part of the eHumanities Desktop in order to test the proposed classifiers.

1 Introduction

In this paper we consider the problem of text categorization by means of *Closed Topic Models* (CTM). Different to *Open Topic Models* (OTM) [1, 2] were topic labels represent content categories which change over time contributed by an open community, e.g. Wikipedia users – topic labels of CTM are defined in advance. Therefore, traditional machine learning techniques such as document classification and clustering techniques can be applied. Most commonly the classification or categorization is based on the Vector Space Model (VSM) [3] – representing textual data with the 'bag of words' (BOW) approach. Despite of its variations this method represents all words within a term-document matrix indexed by a feature weighting measure e.g. term frequency (TF) or inverse document frequency (IDF) or a combination of both. Documents are then judged on basis of

their similarity of term-features (following the idea that similar documents will also share similar features) and can be clustered by e.g. k-means algorithms – where k defines the numbers of predefined categories. Conducting Support Vector Machine (SVM) techniques often better results can be achieved. However, features selection is *the* important part when applying SVM for categorizing a document collection. Despite of linguistic information such as part-of-speech, lemma or stem information, selected features are mainly comprised by content und structure features out of the training corpus. Rising the number of considered features boosts also the complexity in computing the categorization. Therefore, its good performance often applies to a small set of categories or short texts. Using larger documents, a feature reduction technique [4–6] has to be applied. Most often this is done by introducing a certain threshold of the feature weighting function. However, utilized features are still those retained out of the document collection. Following the idea that document categorization is not about the words occurring in the text, but about common concepts texts represent, concept enhancement and feature replacement are the keywords of this paper. Our focus is an alternative representation of individual texts in order to enhance existing classical BOW features and then reduce it to a minimum representation. In recent years a few approaches have been proposed regarding feature enhancement using different resources of knowledge. Bloehdorn, Hortho (2004) [7] proposed a method using background knowledge from an ontology by means of the lexical-semantic net *WordNet* [8], to improve text classification. As one of the first in the field of social network driven methods, Gabrilovich, Markovitch (2005) [9] used directory concepts of the Open Directory Project (ODP), to enhance textual data. Later Gabrilovich, Markovitch (2006) [10] and Zalan Bodo et al. [11] used data from the Wikpedia project for support vector machine based text categorization experiments. Wang, Domeniconi (2008) [12] proposed a semantic kernel technique for text classification also on the basis of the Wikipedia data set. In the field of bio-medicine Xinghua et al. (2006) [13] proposed a Latent Dirichlet Allocation model (LDA), evaluated on the TREC corpus for an enhanced representation of biomedical knowledge. As a commonality, all approaches have utilized the *article title* of Wikipedia to enhance their existing dataset. Evaluation was performed using the famous Reuters and the Movie-Review corpus – comprising the English language. Different to them, in this work we propose a 'knowledge feature breath' on basis of generalized *category concepts* out of a social network. We are labelling individual texts of a document collection with a fixed number of relevant category concept definitions instead of article namespaces. Utilized category information, considered as the key topic information, are subsequently used for a topic generalization (e.g. from topic *tennis* to a more general label *sports*). The idea behind this approach is, that topic related documents share similar generalized topic labels. These predicted labels pose as new, semantically related, features for the categorization process. In contrast to the approaches above, we focus on feature reduction rather than feature extension. Predicted labels act as a substitution of the initial textual feature set. The contribution of this paper is threefold: First, we evaluate the

performance of text classification using Latent Semantic Analysis (LSA) [14], Support Vector Machine (SVM) [15], and a feature-reduced SVM implementation, tested on a large corpus of 29,086 texts comprising 30 different categories. Second, we examine the effect in text classification using the proposed semantic-feature-replacement technique by means of topic generalization. Third, an online categorizer will be introduced that aims to combine a convenient user interface with a framework which is open to arbitrary categorization approaches.

2 Method

Taking a 'knowledge feature breath' in order to extend or reduce an existing document representation with topic-related features, an external knowledge repository is needed. In this context we are utilizing the social ontology of the online encyclopedia *Wikipedia* as a source of terminological knowledge. Concepts are reflected through *Wikipedia articles* and more importantly their corresponding *Wikipedia category* information (Section 2.1). In particular, we make use of the category taxonomy in order to predict generalized topic labels, constructing additional semantically related feature concepts (Section 2.2). Predicted concepts are then subsequently used for the classification task either as feature enhancement or replacement candidates (Section 2.3).

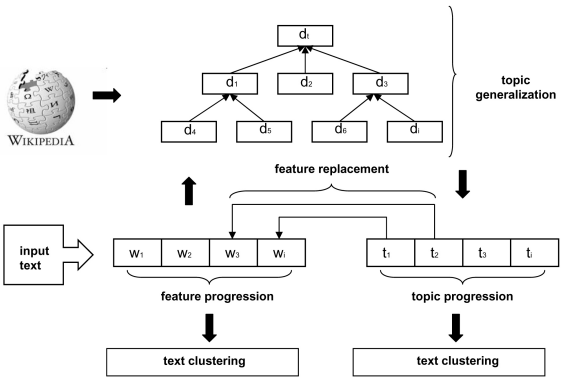


Fig. 1. Text categorization by means of generalized topic concepts.

2.1 Concept Generation

The method of identifying category labels out of the Wikipeda, utilizes the article collection of the social network. Generally speaking, we first try to identify the most adequate article concepts for a given text fragment, and then use associated category information to predict more general topic labels (see Figure

1). The approach of mapping a given text fragment onto the Wikipedia article collection is done following Gabrilovich and Markovitch (2007) [16], by building an inverted vector index. A detailed description of the used minimized representation of the *German Wikipedia* dataset and the method in aligning a given text fragment onto the article collection can be found in Waltinger and Mehler (2009) [1]. At large, we merely parse the entire Wikipedia article collection, and perform a tokenization and lemmatization. Each article and its corresponding lemmata are stored in a vector representation. Therefore, vector entries represent lemmata that occur in the respective article. Each lemma feature is weighted by the TF-IDF scheme [17], which reflects the association or affinity to the corresponding article concepts. In a next step, we invert this vector, defined as V_{art}, using lemmata as the index, and article namespaces as the index entries. The reduction of the vector representation is done by sorting all article concepts (the vector entries) on basis of their affinity scores in descending order and remove those articles concepts whose affinity score is less than five percent of the highest feature weight. Having V_{art} given, we can apply a standard text similarity algorithm, using the cosine metric, in order to identify relevant Wikipedia articles for a given text fragment. In order to access the topic-related category information, we follow Waltinger, Mehler, Heyer (2008) [18], using the assigned article-category hyperlinks within each Wikipedia article page. Thus, the second vector representation, defined as V_{cat}, stores for each article entry its corresponding category concepts. Feature are also weighted through the TF-IDF scheme. Following this, we are able to retrieve the weighted number of unique category concepts K for a given text fragment by iterating over V_{art} and collecting $k_j \in V_{cat}$. Since both vectors V_{art} and V_{cat} are sorted in descending order, the first entries of our vectors correspond to those concepts which fits best to a given input text on basis of our concept vector representation.

2.2 Topic Generalization

Being able to request the most relevant Wikipedia articles – V_{art} – and their corresponding category information – V_{cat} – for a given text fragment, we are following Waltinger, Mehler (2009) [1] in computing a topic generalization. The task to generalize certain topics is defined as making generalizations from specific concepts to a broader context. For example from the term *BASKETBALL* to the more general concept *SPORT* or from *DELEGATE* to *POLITICS*. Again, we are utilizing the category taxonomy of the Wikipedia for this task. The category taxonomy has been extracted in a top-down manner, from the root category namespace: *Category:Contents*, connecting all subordinated categories to its superordinate concepts. We therefore forced the taxonomy into the representation of a directed tree defined as D. Any given text fragment is first mapped to the most specific category concepts as an entry point – $V_{art} \mapsto V_{cat}$ – and then tracked upwardly moving along the taxonomy. Each category concept we meet on the way up inherits the feature weight of the initial category. Therefore, for each edge we have passed a more general concept is derived – comprising the desired topic generalization vector V_{topic}. V_{topic} is also sorted in descending order,

from the most general topics at the beginning to the most specific concepts at the end. See Table 1 and Table 2 for an example of the topic generalization for different domains.

DAS GRÖSSTE KURSPLUS seit 1985 wurde an den acht hiesigen Börsen im vergangenen Jahr erzielt. Beispielsweise zog der Deutsche Aktienindex um 47 Prozent an (vgl. SZ Nr. 302). Trotz Rezession und Hiobsbotschaften von der Unternehmensfront hatten sich zunächst britische und amerikanische Fondsverwalter bei hiesigen Standardwerten engagiert, woraufhin in der zweiten Hälfte des vergangenen Jahres der SZ-Index um 31 Prozent hochgeschnellt war. ...

.

.

Related Articles	Generalized Topics
1. Anlageklasse	1. Finanzierung
2. Bundesanleihe	2. Finanzmarkt
3. Nebenwert	3. Ökonomischer Markt
4. Bullen- und Bärenmarkt	4. Wirtschaft
5. Börsensegment	5. Rechnungswesen

Table 1. Top-5 article and generalized topic concepts for closed topic *stock market*.

Berwerbungsfrist läuft ab: Bis zum 15. Januar müssen die Bewerbungen für die zulassungsbeschränkten Studienplätze bei der Zentralstelle für die Vergabe von Studienplätzen (ZVS) in Dortmund eingetroffen sein. Die notwendigen Unterlagen sind bei den örtlichen Arbeitsämtern, Universitäten ... Weniger Habilitationen: 1992 wurden an den Hochschulen in Deutschland rund 1300 Habilitationsverfahren ...

.

.

Related Articles	Generalized Topics
1. Provadis School of IMT	1. Bildung
2. Approbationsordnung	2. Deutschland
3. Private Hochschule	3. Bildung nach Staat
4. Hochschulabschluss	4. Akademische Bildung
5. Hochschule Merseburg	5. Wissenschaft

Table 2. Top-5 article and generalized topic concepts for closed topic *campus*.

2.3 Feature Replacement

The main focus of this paper is the task of feature replacement for text categorization. In a broader context we address the problem of high dimensionality of BOW approaches due to the amount of comprised features. Feature reduction

techniques contribute to the removal of noise and lower the overfitting in a classification process. In order to judge the importance of comprised features to categories, a weighting function is needed. We make use of the well-known TF-IDF weighting function (see Equation 1), which measures the importance of a feature t_i to the actual document d_j in connection to the entire corpus size N. Therefore an input document d is represented as a data vector $\mathbf{v}_d = [w_{1,d}, w_{2,d}, \ldots, w_{N,d}]^T$, where d_i is defined as a set of features.

$$w_{ij} = tf_{ij} \cdot idf_i = \frac{freq_{ij}}{max_l(freq_{lj})} \cdot \log \frac{N}{n_i} \tag{1}$$

Once having all features weighted, we conduct the concept construction on the basis of the topic generalization proposed in the previous section. In special, for each text we generate twenty topic-related concepts. The ten best article concepts and the ten best category concepts. Note that each concept is tokenized and weighted by w_{ij}. This is done in order to gain affinity scores for different written concepts. Consider for example the category concept: *ACADEMIC EDUCATION*. We resolve this multi-word concept into two individual concepts: *ACADEMIC* and *EDUCATION*. All resolved weighted features are then put to our semantic feature vector defined as V_{sem}. Feature reduction in first place is performed by replacing initial features of the data vector d with lower w_{ij} by corresponding items in V_{sem}. The main idea behind this approach is that related documents share related generalized concepts.

3 Empirical Evaluation

3.1 SVM Settings

Since we were interested in the performance of feature replacement, we conducted different experiments by varying the initial amount of features. First, we used all textual noun, verb and adjective features (C-SVM). Second, we reduced the initial features to noun only features and limited by a threshold (R-SVM). Third, we added all features of V_{sem} to the reduced feature representation (G-SVM). Forth, we replaced the amount of features by the size of V_{sem} (M-SVM). Fifth, we used only the features gained from the topic generalization (V_{sem}) (GO-SVM). Sixth, we used only the features of V_{sem} and additionally reduced it with a certain threshold (MGO-SVM). For the actual text categorization we make use of the kernel-based classification algorithm of the support vector machine (SVM) implementation SVMlight[15] version 6.02. We used SVMs because their good performance in the task of text categorization. For each class an SVM classifier was trained using the linear kernel. The results were evaluated using the leave-one-out cross-validation estimation of SVMlight. For comparison to non-SVM approaches we computed various supervised and unsupervised baselines. First, a random clustering of all documents was performed. Second we conducted a Latent Semantic Analysis (LSA) [19]. It is a dimensionality reduction technique based on singular value decomposition (SVD). The SVD is computed keeping

the k best eigenvalues. In our experiments we defined k as *300*. The resultant matrix was then used for the categorization experiments conducting different clustering techniques including k-means, hierarchical, average linking clustering.

3.2 Vector Settings

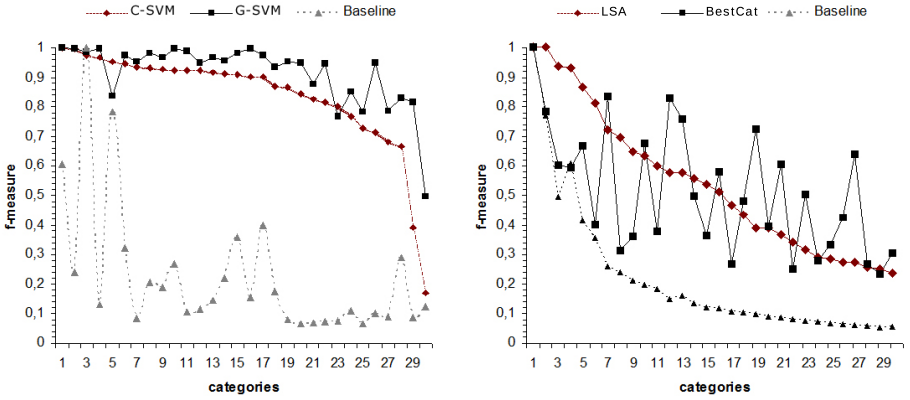
The calculation of our feature vector representation is based upon the German version of Wikipedia (February 2009). After parsing the XML dump comprising 756,444 articles we conducted the preprocessing by lemmatizing all input tokens and removing smaller concepts. We ignored those articles having fewer than five incoming and outgoing links and fewer than 100 non stopwords. The final vector representation comprised 248,106 articles and 620,502 lemmata. The category tree representation consisted of 55,707 category entries utilizing 128,131 directed hyponomy edges.

3.3 Evaluation Corpus

The evaluation of the proposed methods was done by using a large corpus of newspaper articles. We used data of the German newspaper Süddeutsche Zeitung (SZ). The initial corpus comprised 135,546 texts within 96 categories. Due to its unbalanced category-text proportions, an adjusted subset was extracted consisting of 29,086 text, 30 categories and 232,270 unique textual features.

4 Results

The aim of our experiment was to determine the effect of topic generalization in a SVM text classification environment. To which level does feature enhancement by means of semantic concepts boost the classical SVM categorization? To which level can we reduce the initial features set in order to still retain an acceptable performance? How doe unsupervised methods perform compared to supervised? As Figure 4 shows, all SVM (supervised) methods clearly outperform the unsupervised clustering results. With an average F-measure of 0.631, clustering conducting a LSA (see Table 3) obviously performs better than baseline approaches (F-measure of 0.15), but also confirms that for classical text categorization SVMs are most appropriate. Comparing the different SVM implementations (Table 4), we can identify that feature enhancement (G-SVM with an average F-measure of 0.424) boosts with a difference of up to 0.700 (at category *gesp*) the classical SVM implementation using all noun, verb and adjective features (0.778). Yet, in comparison to a much reduced SVM implementation (R-SVM: 0.914) – using only nouns and limited to 5,000 features overall - only minor enhancement can be identified. But looking more closely at the data, we can observe that within categories which perform lower than an F-Measure of 0.900 in the reduced version the G-SVM improves the results. Nevertheless, since the average results of the R-SVM implementation are a priori very high not much improvement could have been expected. Much more interesting is the



(a) (b)

Fig. 2. a) F-Measure results of supervised classification comparing baseline, classical and topic generalized enhanced SVM implementations. b) F-Measure results of unsupervised classification comparing baseline, average linking and best category stream clustering.

aspect of using only the topic generalization concepts for the classification, and discarding all actual features of the text. Using only twenty concepts per text, we still reach a promising F-measure of 0.884 (GO-SVM[1]). Reducing this set to the MGO-SVM[2] implementation, that is only 1000 features for all 29,086 texts the average results show also a very promising 0.855. When comparing thereby the number of used features (see Table 4) using the GO-SVM and the R-SVM, we can identify that we were able to reduce the actual reduced features additional with an average percentage of 80.10%. Therefore, the results for using only the topic generalization concepts for text categorization seem to be very up-and-coming.

5 Online Categorization Using eHumanities Desktop

We have presented an approach to text classification by incorporating social ontologies and showed an evaluation with different settings. In order to put interested users in a position to test the classifiers on their own, we have developed an easy-to-use web interface as part of the eHumanities Desktop [20, 21]. The eHumanities Desktop is an online system for linguistic corpus management, processing and analysis. Based on a well-founded data model to manage resources,

[1] The GO-SVM results are based upon eleven categories, since classification process was still running by the end of the cfp-deadline.

[2] The MGO-SVM was reported on the basis of five computed categories, since the SVM was still computing by the end of the cfp-deadline.

Implementation	F-Measure
G-SVM	0.915
R-SVM	0.914
M-SVM	0.913
C-SVM	0.836
GO-SVM	0.884[1]
MGO-SVM	0.855[2]
LSA	0.631
Random	0.15

Table 3. Average F-Measure results of the entire classification experiments. C-SVM refers to the SVM implementation using all textual features; G-SVM to a feature enhanced, R-SVM and M-SVM to the feature reduced implementations. GO-SVM and MGO-SVM refer to the experiments using only topic-related concepts as the feature set.

users and groups it offers application modules to perform tasks of text preprocessing, information retrieval and linguistic analysis. The set of functionality is easily extensible by new application modules - as for example the *Categorizer*. The *Categorizer* aims to combine a convenient user interface with a framework which is open to arbitrary categorization approaches. The typical user does not want to bother with details of a given method and what preprocessing needs to be done. Using the *Categorizer* all that needs to be done is to pick a classifier, specify the input document (e.g. plain text, html or pdf) and start the categorization. Alternatively it is also possible to enter the input text directly via cut&paste. The classifiers themselves are defined in terms of a *eHumanities Desktop Classifier Description*, an XML based language to specify how to connect a given input document to a classifier. The language is kept as flexible as possible to be open to arbitrary algorithms. In case of a SVM a classifier description defines, among others, which kind of text preprocessing needs to be done, what features and which models to use and how an implementation of the SVM needs to be called. Integrating a new classification algorithm usually does not take more than putting the program on the server and writing a proper classifier description. Please note that the category models are 'normal' documents of the corpus management system as are the classifier descriptions. Thus they can easily be shared among users. Figure 5 shows an example of how the *Categorizer* can be used to categorize text. The content has directly been inserted via cut&paste from the online portal of a German newspaper. The classifier is SVM-based and trained on categories of the *Süddeutsche Zeitung*. The table shows the results of the categorization with the best performing category at the top.

6 Conclusions

This paper presented a study on text categorization by means of closed topic models. We proposed a SVM based approach using a semantic feature replace-

ID	Name	No.T.	Feat.Reduc.	C-SVM	R-SVM	G-SVM	Imp. 1	Imp. 2
1	baro	465	-87.00%	0.995	0.998	0.998	+0.000	+0.003
2	camp	345	-87.33%	0.913	0.953	0.956	+0.003	+0.043
3	diew	276	-84.96%	0.925	0.994	0.995	+0.001	+0.070
4	fahr	313	-88.31%	0.924	0.978	0.989	+0.011	+0.065
5	film	2457	-87.57%	0.865	0.955	0.954	-0.001	+0.089
6	fird	2213	-71.87%	0.902	0.971	0.973	+0.002	+0.071
7	firm	1339	-83.59%	0.969	0.995	0.995	+0.000	+0.026
8	gesp	1234	-91.08%	0.393	0.751	0.817	+0.066	+0.424
9	inha	1933	-55.11%	0.974	0.988	0.987	-0.001	+0.013
10	kost	533	-90.57%	0.926	0.964	0.966	+0.002	+0.040
11	leut	911	-78.69%	0.903	0.994	0.996	+0.002	+0.093
12	loka	1953	-76.94%	0.728	0.772	0.781	+0.009	+0.053
13	mein	2240	-79.82%	0.923	0.961	0.951	-0.010	+0.028
14	mitt	677	-74.51%	0.171	0.485	0.495	+0.010	+0.324
15	nchg	1105	-84.99%	0.799	0.761	0.769	+0.008	-0.020
16	nrwk	349	-76.17%	0.871	0.957	0.936	-0.021	+0.065
17	nrwp	297	-85.21%	0.952	0.846	0.838	-0.008	-0.114
18	nrww	342	-76.46%	0.932	0.983	0.983	+0.000	+0.051
19	reit	286	-80.09%	0.933	0.946	0.953	+0.007	+0.020
20	schf	542	-61.92%	0.682	0.795	0.785	-0.010	+0.103
21	spek	375	-68.73%	0.712	0.975	0.950	-0.025	+0.238
22	spfi	318	-90.84%	0.908	0.984	0.983	-0.001	+0.075
23	stdt	700	-75.89%	0.767	0.850	0.853	+0.003	+0.086
24	szen	2314	-82.66%	0.827	0.869	0.878	+0.009	+0.051
25	szti	336	-56.82%	0.947	0.979	0.973	-0.006	+0.026
26	thkr	1613	-90.24%	0.817	0.939	0.945	+0.006	+0.128
27	tvkr	2355	-78.87%	0.846	0.949	0.951	+0.002	+0.105
28	woch2	375	-88.17%	1.00	1.00	1.00	+0.000	+0.000
29	zwif	409	-84.06%	0.666	0.862	0.829	+0.033	+0.163
30	zwiz	481	-84.69%	0.918	0.968	0.969	+0.001	+0.051

Table 4. Results of SVM-Classification comparing R-SVM and G-SVM (Imp. 1) and C-SVM and G-SVM (Imp. 2). No.T. refers to the number of texts within a category. Feat.Reduc. shows the amount of feature reduction, comparing the initial- and the reduced feature set.

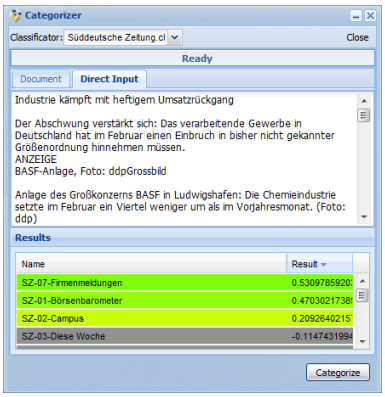


Fig. 3. Screenshot showing the eHumanities Desktop Categorizer.

ment. New features are created on the basis of the social network Wikipedia conducting a topic generalization using category information. Generalized concepts are then used as a replacement of conventional features. We examined different methods in enhancing, replacing and the deletion of features during the classification process. In addition, we offer an easy-to-use web interface as part of the eHumanities Desktop in order to test the proposed classifiers.

Acknowledgment

We gratefully acknowledge financial support of the German Research Foundation (DFG) through the EC 277 *Cognitive Interaction Technology*, the SFB 673 *Alignment in Communication* (X1), the Research Group 437 *Text Technological Information Modeling*, the DFG-LIS-Project *P2P-Agents for Thematic Structuring and Search Optimization in Digital Libraries* and the *Linguisitc Networks* project funded by the German Federal Ministry of Education and Research (BMBF) at Bielefeld University.

References

1. Waltinger, U., Mehler, A.: Social semantics and its evaluation by means of semantic relatedness and open topic models. In: Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence. (2009)
2. Mehler, A., Waltinger, U.: Enhancing document modeling by means of open topic models: Crossing the frontier of classification schemes in digital libraries by example of the DDC. Appears in Library Hi Tech (2009)
3. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley, Reading, Massachusetts (1989)
4. Taira, H., Haruno, M.: Feature selection in svm text categorization. In: AAAI '99/IAAI '99: Proceedings of the 6.th national conference on AI, Menlo Park, CA, USA, American Association for Artificial Intelligence (1999) 480–486

5. Kim, H., Howland, P., Park, H.: Dimension reduction in text classification with support vector machines. J. Mach. Learn. Res. **6** (2005) 37–53
6. Dasgupta, A., Drineas, P., Harb, B., Josifovski, V., Mahoney, M.W.: Feature selection methods for text classification. In: KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2007) 230–239
7. Andreas, S.B., Hotho, A.: Boosting for text classification with semantic features. In: In Proceedings of the MSW 2004 Workshop at the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. (2004) 70–87
8. Fellbaum, C., ed.: WordNet. An Electronic Lexical Database. The MIT Press (1998)
9. Gabrilovich, E., Markovitch, S.: Feature generation for text categorization using world knowledge. In: Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence, Edinburgh, Scotland (2005) 1048–1053
10. Gabrilovich, Markovitch: Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. Proceedings of the Twenty-First National Conference on Artificial Intelligence, Boston, MA (2006)
11. Zalan Bodo, Zsolt Minier, L.C.: Text categorization experiments using wikipedia. (2007) 66–72
12. Wang, P., Domeniconi, C.: Building semantic kernels for text classification using wikipedia. In: KDD08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2008) 713–721
13. Xinghua, L., Bin, Z., Atulya, V., ChengXiang, Z.: Enhancing text categorization with semantic-enriched representation and training data augmentation. (2006)
14. Landauer, T., Dumais, S.: A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review **104**(1) (1997) 211–240
15. Joachims, T.: Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers, Norwell, MA, USA (2002)
16. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. Proceedings of the 20th International Joint Conference on Artificial Intelligence (2007) 6–12
17. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
18. Waltinger, U., Mehler, A., Heyer, G.: Towards automatic content tagging: Enhanced web services in digital libraries using lexical chaining. In: 4rd International Conference on Web Information Systems and Technologies (WEBIST '08), 4-7 May, Funchal, Portugal, Barcelona (2008)
19. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science **41**(6) (1990) 391–407
20. Mehler, A., Gleim, R., Waltinger, U., Ernst, A., Esch, D., Feith, T.: ehumanities desktop — eine webbasierte arbeitsumgebung für die geisteswissenschaftliche fachinformatik. In: Proceedings of the Symposium "Sprachtechnologie und eHumanities", 26.–27. Februar, Duisburg-Essen University. (2009)
21. Gleim, R., Waltinger, U., Ernst, A., Mehler, A., Feith, T., Esch, D.: eHumanities Desktop - an online system for corpus management and analysis in support of computing in the humanities. In: Proceedings of the Demonstrations Session at EACL 2009, Athens, Greece, Association for Computational Linguistics (April 2009) 21–24