Syntax versus Semantics:
Analysis of Enriched Vector Space Models

Benno Stein[1] and **Sven Meyer zu Eissen**[1] and **Martin Potthast**[2]

Abstract. This paper presents a robust method for the construction of collection-specific document models. These document models are variants of the well-known vector space model, which relies on a process of selecting, modifying, and weighting index terms with respect to a given document collection. We improve the step of index term selection by applying statistical methods for concept identification. This approach is particularly suited for post-retrieval categorization and retrieval tasks in closed collections, which is typical for intranet search.

We compare our approach to "enriched" vector-space-based document models that employ knowledge of the underlying language in the form of external semantic concepts. Primary objective is to quantify the impact of a purely syntactic analysis in contrast to a semantic enrichment in the index construction step. As a by-product we provide an efficient and language-independent means for vector space model construction, whereas the resulting document models perform better than the standard vector space model.

Keywords vector space model, concept identification, semantic concepts, text categorization, evaluation measures

1 INTRODUCTION

Each text retrieval task that is automated by a computer relies on some kind of document model, which is an abstraction of the original document d. The document model must be tailored well with respect to the retrieval task in question: It determines the quality of the analysis, and—diametrically opposed—its computational complexity. Though its obvious simplicity the vector space model has shown great success in many text retrieval tasks [11; 12; 16; 15], and, the analysis of this paper uses this model as its starting point.

The standard vector space model abstracts a document d toward a vector $\mathbf{d}$ of weighted index terms. Each term t that is included in $\mathbf{d}$ derives from a term $\tau \in d$ by affix removal, which is necessary to map morphological variants of τ onto the same stem t. The respective term weights in $\mathbf{d}$ account for the different discriminative power of the original terms in d and are computed according to some frequency scheme. The main application of the vector space model is document similarity computation.

In this paper we focus on the index construction step and, in particular, on index term selection. Other concepts of the vector space model, such as the term weighting scheme or its disregard of word order are adopted.

[1] Faculty of Media, Media Systems.
Bauhaus University Weimar, 99421 Weimar, Germany
{benno.stein | sven.meyer-zu-eissen}@medien.uni-weimar.de
[2] Faculty of Computer Science.
Paderborn University, 33098 Paderborn, Germany

1.1 A Note on Semantics

We classify an index construction method as being semantic if it relies on additional domain knowledge, or if it exploits external information sources by means of some inference procedure, or both. Short documents may be similar to each other from the (semantic) viewpoint of a human reader, while the related instances of the vector space model do not reflect this fact because of the different words used. Index term enrichment can account for this by adding synonymous terms, hypernyms, hyponyms, or co-occurring terms [7].

Semantic approaches are oriented at the human understanding of language and text, and, as given in the case of ontological index term enrichment, they are computationally efficient. However, the application of the semantic approaches is problematic, if, for instance, the document language is unknown or if a document combines passages from several languages. Moreover, there are situations where semantic approaches can even impair the retrieval quality: Consider a document collection with specialized texts, then ontological index term enrichment will move the specific character of a text toward a more general understanding. As a consequence, the similarity of highly specialized text is diluted in favor of less specialized text—which compares to the effect of adding noise.

1.2 Contributions

We investigate variants of the vector space model with respect to their classification performance. Starting point is the standard vector space model where the step of index term selection is improved by a syntactic approach for concept identification; the resulting model is compared to semantically enriched vector space models. The syntactic concept identification approach is based on a collection-specific suffix tree analysis. In a nutshell, the paper's underlying question may be summarized as follows:

Can syntactically determined concepts keep up with a semantically motivated index term enrichment?

To answer this question we have set up a number of text categorization experiments with different clustering algorithms. Since these algorithms are susceptible to various side effects, we will also present results that rely on an objective similarity assessment statistic: the measure of expected density, $\overline{\rho}$. Perhaps the most interesting result may be anticipated: The positive effect of semantic index term enrichment, which has been reported by some authors in the past, could hardly be observed in our comprehensive analysis.

The remainder of the paper is organized as follows. Section 2 presents a taxonomy of index construction methods and outlines commonly used technology, and Section 3 reports on similarity analysis and unsupervised classification experiments.

2 INDEX CONSTRUCTION FOR DOCUMENT MODELS

This section organizes the current practice of index construction for vector space models. In particular, we review the concept of a document model and propose a classification scheme for both popular and specialized index construction principles.

A document d can be viewed under different aspects: layout, structural or logical setup, and semantics. A computer representation $\mathbf{d}$ of d may capture different portions of theses aspects. Note that $\mathbf{d}$ is designed purposefully, with respect to the structure of a formalized query, $\mathbf{q}$, and also with having a particular retrieval model in mind. A retrieval model, $\mathcal{R}$, provides the linguistic rationale for the model formation process behind the mapping $d \mapsto \mathbf{d}$. This mapping involves an inevitable simplification of d that should be

1. quantifiable,
2. useful with respect to the information need, and
3. tailored to $\mathbf{q}$, the formalized query.

The retrieval model $\mathcal{R}$ gives answers to these points, be it theoretically or empirically, and provides a concrete means, $\rho(\mathbf{q}, \mathbf{d})$, for quantifying the relevance between a formalized query $\mathbf{q}$ and a document's computer representation $\mathbf{d}$. Note that $\rho(\mathbf{q}, \mathbf{d})$ is often specified in the form of a similarity measure φ.

Together, the computer representation $\mathbf{d}$ along with the underlying retrieval model $\mathcal{R}$ form the document model; Figure 2 illustrates the connections.

Let D be a document collection and let T be the set of all terms that occur in D. The vector space model $\mathbf{d}$ of a document d is a vector of $|T|$ weights, each of which quantifying the "importance" of some index term in T with respect to d.[3] This quantification must be seen against the background that one is interested in a similarity function φ that maps from the vectors $\mathbf{d_1}$ and $\mathbf{d_2}$ of two documents d_1, d_2 into the interval $[0; 1]$ and that has the following property: If $\varphi(\mathbf{d_1}, \mathbf{d_2})$ is close to 1 then the documents d_1 and d_2 are similar; likewise, a value close to zero indicates a high dissimilarity. Note that document models and similarity functions determine each other: The vector space model and its variants are amenable to the cosine similarity (= normalized dot product) in first place, but can also be used in connection with Euclidean distance, overlap measures, or other distance concepts.

Under the vector space paradigm the document model construction process is determined in two dimensions: index construction and weight computation. In the following we will concentrate on the former dimension since this paper contributes right here. We have classified the index construction principles for vector space models in four main classes, which are shown in Figure 1.

[3] Note that, in effect, the vector space model is a computer representation of a the textual content of a document. However, in the literature the term "vector space model" is also understood as a retrieval model with a certain kind of relevance computation.

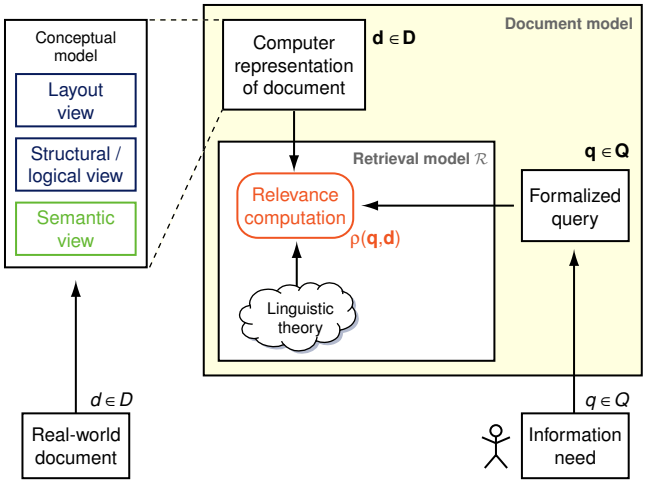

Figure 2. In the end, an information need q is satisfied by a real-world document d. A computer support for this retrieval task requires an abstraction of q and d towards $\mathbf{q}$ and $\mathbf{d}$. The rationale for this abstraction comes from a linguistic theory and is operationalized by a retrieval model $\mathcal{R}$.

Index Term Selection. Selection methods further divide into inclusion and exclusion methods. An important exclusion method is stopword removal: Common words, such as prepositions or conjunctions, introduce noise and provide no discriminating similarity information; they are usually discarded from the index set. However, there are special purpose models (e. g. for text genre identification) that rely on stopword features [13; 9].

The standard vector space model does not apply an inclusion method but simply takes the entire set T without stopwords. More advanced vector space models use also n-grams, i. e., continuous sequences of n words, $n \leq 4$, which occur in the documents of D. Since the usage of n-grams entails the risk of introducing noise, not all n-grams should be added but threshold-based selection methods be applied, which rely on the information gain or a similar statistic [6].

Index Term Modification. Most term modification methods aim at generalization. A common problem in this connection is the mapping of morphologically different words that embody the same concept onto the same index term. So-called stemming algorithms apply here; their goal is to find canonical forms for inflected or derived words, e. g. for declined nouns or conjugated verbs. Since the "unification" of words with respect to gender, number, time, and case is a language-specific issue, rule-based stemming algorithms require the development of specialized rule sets for each language. Recall that

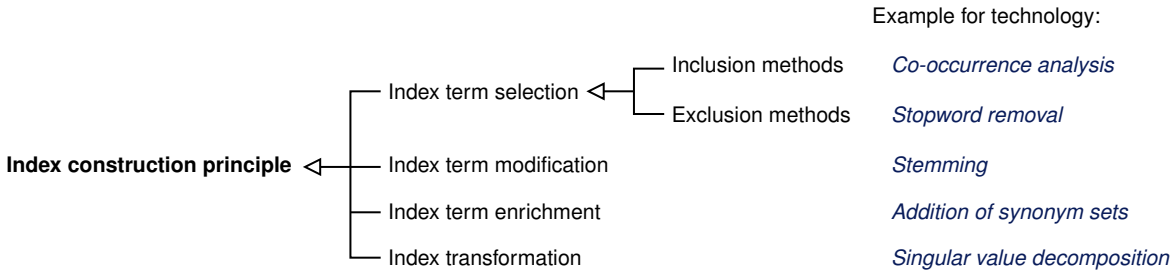

Figure 1. A taxonomy of index construction principles for vector space models.

the application of language-specific rule sets requires the problem of language detection both in unilingual and multilingual documents to be solved.

Index Term Enrichment. We classify a method as term enriching, if it introduces terms *not* found in T. By nature, meaningful index term enrichment must be semantically motivated and exploit linguistic knowledge. A standard approach is the—possibly transitive—extension of T by synonyms, hypernyms, hyponyms, and co-occurring terms. The extension shall alleviate the problem of different writing styles, or of vocabulary variations observed in very small document snippets as they are returned from search engines.

Note that these methods are not employed to address the problem of polysemy, since the required in-depth analysis of the term context is computationally too expensive for many similarity search applications.

Index Transformation. In contrast to the construction methods mentioned before, transformation methods operate on all document vectors of a collection D at the same time by analyzing the term-document matrix, A. A popular index transformation method is latent semantic indexing (LSI), which uses a singular value decomposition of A in order to improve query rankings and similarity computations [2; 1; 8]. For this purpose, the document vectors are projected into a low-dimensional space that is spanned by the eigenvectors that belong to the largest singular values of the decomposition of A.

2.1 Discussion

Index terms that consist of a single word can be found by a skillful analysis of prefix frequency and prefix length. This idea can be extended to the identification of compound word concepts in written text. If continuous sequences of n words occur significantly often, then it is likely that these words form a concept. Put another way, concept detection reduces to the identification of frequent n-grams.

n-grams as a replacement for index term enrichment has been analyzed by several authors in the past, with moderate success only [6]. We explain the disappointing results with noise effects, which dominate the positive impact of few additional concepts: Most authors apply a strategy of "complete extension"; i. e., they add all 2-grams and 3-grams to the index vector. However, when analyzing the frequency distribution of n-grams, it becomes clear that only a small fraction of all compound word sequences is statistically relevant.

The advantages of syntactical (statistical) methods for index construction can be summarized as follows:

1. language independence
2. robustness with respect to multi-lingual documents
3. tailored indexes for retrieval tasks on closed collections

An obvious disadvantage may be the necessary statistical mass: Syntactical index construction cannot work if only few, very small document snippets are involved. This problem is also investigated in the next section, where the development of the index quality is compared against the underlying collection size.

As an aside, statistical stemming and the detection of compound word concepts are essentially the same—the level of granularity makes the difference: Stemming means frequency analysis at the level of characters; likewise, the identification of concepts means frequency analysis at the level of words.

3 ANALYSIS OF ENRICHED VECTOR SPACE MODELS

Existing reports on the impact of index term selection and index term enrichment are contradictory [4; 5; 7], and not all of the published performance improvements could be reproduced [6]. Most of this research analyzes the effects of a modified vector space model on typical information retrieval tasks, such as document clustering or query answering.

Note that clustering results that have been obtained by employing the same cluster algorithm under different document models may tell us two things: (*i*) whether one document model captures more of the "gist" of the original document d than another model, and, (*ii*) whether the cluster algorithm is able to take advantage of this added value.

A cluster algorithm's performance depends on various parameters, such as the cluster number, its randomized start configuration, or preset similarity thresholds, etc., which renders a comparison difficult. Moreover, there is the prevalently observed effect that different cluster algorithms behave differently sensitive to document model "improvements". From an analysis point of view the following questions arise:

1. Which cluster algorithm shall define the baseline for a comparison (the best for the dataset, the most commonly used, the simplest)?
2. Given several clustering results obtained by the same cluster algorithm, which result can be regarded as meaningful (the best, the worst, the average)?

Especially to the second point less attention is paid in current research: Common practice is to select the best result compared to a given reference classification, e. g. by maximizing the F-Measure value—ignoring that such a combined usage of unsupervised/supervised methods is far away from reality.[4]

An objective way to rank different document models is to compare their ability to *capture the intrinsic similarity relations* of a given collection D. Basic idea is the construction of a similarity graph, measuring its conformance to a reference classification, and analyzing the improvement or decline of this conformance under some document model. Exactly this is operationalized in form of the $\overline{\rho}$-measure that is introduced below; it enables one to evaluate differences in the similarity concepts of alternative document models without being dependent on a cluster algorithm.[5]

Hence, the performance analyses presented in this section comprise two types of analyses: (*i*) Experiments that, based on $\overline{\rho}$, quantify objective improvements or declines of a document model, (*ii*) experiments that, based on the F-Measure, quantify the effects of a document model onto different cluster algorithms.

3.1 A Measure of Expected Density: $\overline{\rho}$

As before let $D = \{d_1, \ldots, d_n\}$ be a document collection whose corresponding computer representations are denoted as $\mathbf{d}_1, \ldots, \mathbf{d}_n$. A similarity graph $G = \langle V, E, \varphi \rangle$ for D is a graph where a node in V represents a document and an edge $(d_i, d_j) \in E$ is weighted with the similarity $\varphi(\mathbf{d}_i, \mathbf{d}_j)$.

A graph $G = \langle V, E, w \rangle$ is called sparse if $|E| = \mathcal{O}(|V|)$; it is called dense if $|E| = \mathcal{O}(|V|^2)$. Put another way, we can compute the density θ of a graph from the equation $|E| = |V|^\theta$. With

[4] This issue is addressed in [14].

[5] The $\overline{\rho}$-measure was originally introduced in [14], as an alternative for the Davies-Bouldin-Index and the Dunn-Index, in order to evaluate the quality of cluster algorithms for text retrieval applications.

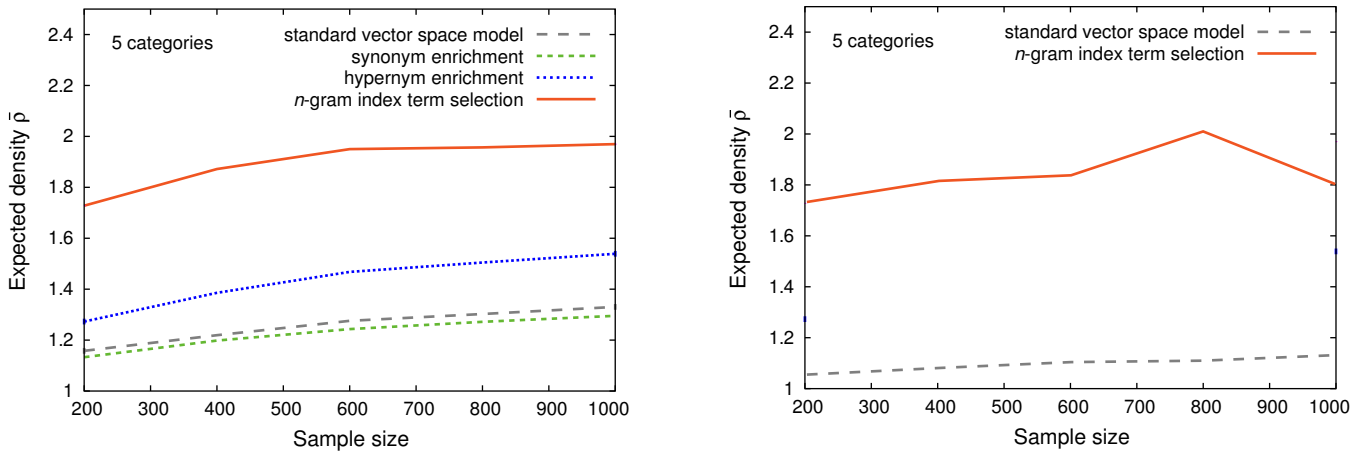

Figure 3. Comparison of the standard vector space model, two semantically enriched models (synonym, hypernym), and a vector space model with syntactically identified concepts (n-gram) in two languages: The left-hand graph illustrates the development of $\overline{\rho}$ depending on the collection size for an English document collection; the right-hand graph compares the n-gram vector space model to the standard model for a German document collection.

$w(G) := |V| + \sum_{e \in E} w(e)$, this relation extends naturally to weighted graphs:[6]

$$w(G) = |V|^{\theta} \quad \Leftrightarrow \quad \theta = \frac{\ln\big(w(G)\big)}{\ln\big(|V|\big)}$$

Obviously, θ can be used to compare the density of each induced subgraph $G' = \langle V', E', w' \rangle$ of G to the density of G: G' is sparse (dense) compared to G if the quotient $w(G')/(|V'|^{\theta})$ is smaller (larger) than 1. This consideration provides a key to quantify a document model's ability to capture the intrinsic similarity relations of G, and hence, of the underlying collection.

Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be an exclusive categorization of D in k distinct categories, that is to say, $C_i, C_j \subseteq D$ with $C_i \cap C_j = \emptyset$ and $\cup_{i=1}^{k} C_i = D$, and let $G_i = \langle V_i, E_i, \varphi \rangle$ be the induced subgraph of G with respect to category C_i. Then the expected density of $\mathcal{C}$ is defined as follows.

$$\overline{\rho}(\mathcal{C}) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^{\theta}}, \quad \text{where} \quad |V|^{\theta} = w(G)$$

Since the edge weights resemble the similarity of the documents associated with V, a higher value of $\overline{\rho}$ indicates a better modeling of a collection's similarity relations.

3.2 Syntax versus Semantics: Variants of the Vector Space Model

Aside from the standard vector space model our analysis compares the following three vector space model variants:

1. *Syntactic Term Selection.* Within this variant the index term selection step also considers syntactically identified concepts, i. e., 2-grams, 3-grams, and 4-grams. To identify the significant n-grams the document collection D is inserted into a suffix tree and a statistical successor variety analysis is applied. The operationalized principle behind this analysis is the peak-and-plateau method [5], for which we have developed a refinement in our working group.
2. *Semantic Synonym Enrichment.* Within this variant of semantic term enrichment the so-called synsets from Wordnet for nouns are added [3]; this procedure has been reported to work well for categorization tasks [7]. Note that adding synonyms to all index terms of a document vector will introduce a lot of noise, and hence only the top-ranked 10% of the index terms (respecting the employed term weighting scheme) are selected for enrichment.
3. *Semantic Hypernym Enrichment.* This variant of semantic term enrichment relies also on Wordnet: a sequence of up to four consecutive hypernyms is substituted for each noun. The rationale is as follows. Documents dealing with closely related—but still different—topics often contain terms which derive from a single hypernym representing their common category. The enrichment proposed here yields a stronger similarity between such documents without generalizing too much.

Index term weighting of both unigrams and n-grams follows the $tf \cdot idf$-scheme; stopwords are not indexed and unigram stemming is done according to Porter's algorithm.

Discussion. The resulting graphs in Figure 3 as well as the comparison in Table 1 show that the syntactic approach outperforms both semantic approaches. From the semantic variants only the semantic hypernym enrichment is above the baseline; note that this happens even if a large number synsets is added. We explain the results as follows: Index terms with a high term weight typically belong to a special vocabulary, and, from a semantic point of view, they are used deliberately so that adding their synsets will tend to *decrease* their importance. Likewise, adding the synsets of low-weighted terms has no effect other than adding noise since the importance of these terms will be *increased without a true rationale*.

Vector space model variant	F-min	F-max	F-av.
	(sample size 1000, 10 categories)		
standard vector space model	—baseline—		
synonym enrichment	-8%	+4%	-2%
hypernym enrichment	+5%	+12%	+3%
n-gram index term selection	+15%	+6%	+8%

Table 1. The table shows the improvements of the averaged F-Measure values that were achieved with the cluster algorithms k-means and MajorClust for the investigated variants of the vector space model.

[6] $w(G)$ denotes the total edge weight of G plus the number of nodes, $|V|$, which serves as adjustment term for graphs with edge weights in $[0; 1]$.

3.3 Test Corpus and Sample Formation

Experiments have been conducted with samples from RCV1, a short hand for "Reuters Corpus Volume 1" [10], as well as with documents from German newsgroup postings.

RCV1 is a document collection that was published by the Reuters Corporation for research purposes. It contains more than 800,000 documents each of which consisting of a few hundred up to several thousands words. The documents are tagged with meta information like category (also called topic), geographic region, or industry sector. There are 103 different categories, which are arranged within a hierarchy of the four top level categories "Government, Social", "Economics", "Markets", and "Corporate, Industrial". Each of the top level categories defines the root of a tree of sub-categories, where each child node fine grains the information given by its parent. Note that a document d can be assigned to several categories $c_1, \ldots, c_p$, and that d does also belong to all ancestor categories of some category c_i.

Within our experiments two documents d_i, d_j are considered to belong to the same category if they share the same top level category c_t and the same most specific category c_s. Moreover, the test sets are constructed in such a way that there is no document d_i whose most specific category c_s is an ancestor of the most specific category of some other document d_j.

The samples were formed as follows: For the analysis of the intrinsic similarity relations based on $\overline{p}$, the sample sizes ranged from 200 to 1000 documents taken from 5 categories. For the analysis of the categorization experiments, based on cluster algorithms and evaluated with the F-Measure, the sample sizes were 1000 documents taken from 10 categories.[7]

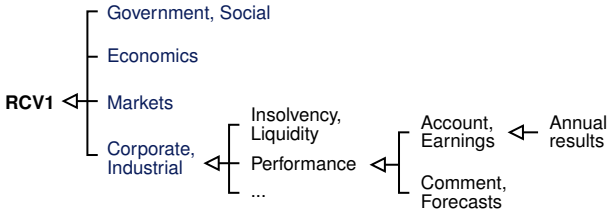

Figure 4. Category organization of the RCV1 corpus showing the four top level categories from which "Corporate, Industrial" is further refined.

4 SUMMARY

This paper provided a comparison of syntactical and semantic methods for the construction of vector space models; the special focus was index term selection. Interestingly, little attention has been paid to the mentioned syntactical methods in connection with text retrieval tasks. Following results of our paper shall be emphasized:

- With syntactically identified concepts significant improvements can be achieved for categorization tasks.
- The benefit of semantic term enrichment is generally overestimated.
- The $\overline{p}$-measure provides an "algorithm-neutral" approach to analyze the similarity knowledge contained in document models.

[7] To make our analysis results reproducible for other researchers, meta information files that describe the compiled test collections have been recorded; they are available upon request.

Note that the last point may be interesting to develop accepted benchmarks to compare research efforts related to document models or similarity measures.

Though syntactical analyses must not be seen as a cure-all for the index construction of vector space models, they provide advantages over semantic methods, such as language independence, robustness, and tailored index sets. With respect to several retrieval tasks they can keep up with semantic methods—however, our results give no room for an over-simplification: Both paradigms have the potential to outperform the other.

References

[1] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien, 'Using Linear Algebra for Intelligent Information Retrieval', Technical Report UT-CS-94-270, Computer Science Department, (dec 1994).

[2] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, 'Indexing by Latent Semantic Analysis', *Journal of the American Society of Information Science*, **41**(6), 391–407, (1990).

[3] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[4] W. B. Frakes, 'Term conflation for information retrieval', in *SIGIR '84: Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 383–389, Swinton, UK, (1984). British Computer Society.

[5] W. B. Frakes and Ricardo Baeza-Yates, *Information retrieval: Data Structures and Algorithms*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1992.

[6] Johannes Fürnkranz, 'A Study Using n-gram Features for Text Categorization', Technical report, Austrian Institute for Artificial Intelligence, (1998). Technical Report OEFAI-TR-9830.

[7] A. Hotho, S. Staab, and G. Stumme, 'Wordnet improves text document clustering', in *Proceedings of the SIGIR Semantic Web Workshop*, (2003).

[8] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala, 'Latent semantic indexing: a probabilistic analysis', in *PODS '98: Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp. 159–168, New York, NY, USA, (1998). ACM Press.

[9] Andreas Rauber and Alexander Müller-Kögler, 'Integrating automatic genre analysis into digital libraries', in *ACM/IEEE Joint Conference on Digital Libraries*, pp. 1–10, (2001).

[10] T.G. Rose, M. Stevenson, and M. Whitehead, 'The Reuters Corpus Volume 1 - From Yesterday's News to Tomorrow's Language Resources', in *Proceedings of the Third International Conference on Language Resources and Evaluation*, (2002).

[11] G. Salton and M. E. Lesk, 'Computer Evaluation of Indexing and Text Processing', *ACM*, **15**(1), 8–36, (January 1968).

[12] Karen Sparck-Jones, 'A statistical interpretation of term specificity and its application in retrieval', *Journal of Documentation*, **28**, 11–21, (1972).

[13] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, 'Text genre detection using common word frequencies', in *Proceedings of the 18th Int. Conference on Computational Linguistics*, Saarbrücken, Germany, (2000).

[14] Benno Stein, Sven Meyer zu Eißen, and Frank Wißbrock, 'On Cluster Validity and the Information Need of Users', in *Proceedings of the 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 03),*

Benalmádena, Spain, ed., M. H. Hanza, pp. 216–221, Anaheim, Calgary, Zurich, (September 2003). ACTA Press.
[15] Michael Steinbach, George Karypis, and Vipin Kumar, 'A comparison of document clustering techniques', Technical Report 00-034, Department of Computer Science and Egineering, University of Minnesota, (2000).
[16] Yiming Yang and Jan O. Pedersen, 'A comparative study on feature selection in text categorization', in *Proceedings of ICML-97, 14th International Conference on Machine Learning*, ed., Douglas H. Fisher, pp. 412–420, Nashville, US, (1997). Morgan Kaufmann Publishers, San Francisco, US.