The ESA Retrieval Model Revisited

Maik Anderka
Faculty of Media, Media Systems
Bauhaus University Weimar
99421 Weimar, Germany
maik.anderka@uni-weimar.de

Benno Stein
Faculty of Media, Media Systems
Bauhaus University Weimar
99421 Weimar, Germany
benno.stein@uni-weimar.de

ABSTRACT

Among the retrieval models that have been proposed in the last years, the ESA model of Gabrilovich and Markovitch received much attention. The authors report on a significant improvement in the retrieval performance, which is explained with the semantic concepts introduced by the document collection underlying ESA. Their explanation appears plausible but our analysis shows that the connections are more involved and that the "concept hypothesis" does not hold. In our contribution we analyze several properties that in fact affect the retrieval performance. Moreover, we introduce a formalization of ESA, which reveals its close connection to existing retrieval models.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Retrieval models*; H.1.1 [Models and Principles]: Systems and Information Theory—*General systems theory*
General Terms: Performance, Theory
Keywords: Explicit Semantic Analysis, Retrieval Models

1. EXPLICIT SEMANTIC ANALYSIS

The Explicit Semantic Analysis (ESA) is a retrieval model proposed by Gabrilovich and Markovitch [4]. It was originally introduced to compute the semantic relatedness of natural language texts, and it yields significant improvements compared to the vector space model or hidden variable models like LSI [3]. Recently, the ESA model was applied in several applications, among others to information retrieval [2], cross-lingual information retrieval [9, 10], text categorization [1, 5], and electronic career guidance [6].

The ESA representation of a real-world document d is a vector $\mathbf{d}_{ESA}$ whose elements are the cosine similarities between d and all documents in a collection D_I, called index collection here. The supposed rationale of the ESA retrieval model is that each document in D_I functions as a semantic concept to which the original document d is compared: $\mathbf{d}_{ESA}$ is understood as a projection of d into the concept space spanned by D_I. The semantic relatedness between two documents d_1 and d_2 is computed by the cosine similarity between the ESA vectors of d_1 and d_2. The authors of [4] attribute the retrieval performance of ESA to the fact that each document in the index collection describes exactly one concept, and that these concepts are "orthogonal". We refer to these properties as *concept hypothesis*. Gabrilovich and Markovitch used Wikipedia as index collection, since it fulfills these requirements because of its "encyclopedic characteristic". However, as the following analysis shows, the concept hypothesis does not hold.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

Table 1: The correlation coefficient achieved with ESA based on different index collections depending on the number of index documents. Bold numbers indicate the row maximum.

Index collection	number of index documents					
	1 000	10 000	50 000	100 000	150 000	200 000
VSM (baseline)	0.717	0.717	0.717	0.717	0.717	0.717
Wikipedia, $tf \cdot idf$	0.742	**0.784**	0.782	0.782	0.781	0.781
Merged Topics, $tf \cdot idf$	0.738	0.767	0.768	0.769	0.769	**0.777**
Reuters, $tf \cdot idf$	0.767	0.795	**0.802**	0.800	0.800	0.800
Wikipedia, tf	0.704	0.724	0.732	0.732	**0.734**	0.732
Random Gaussian, tf	0.703	0.716	**0.717**	0.717	0.717	0.717

2. EMPIRICAL EVALUATION

We use the same experimental set-up as the authors in [4], a test collection consisting of 50 documents from the Australian Broadcasting Corporation's news mail service [7].[1] For this collection human similarity assessments for the 1 225 document relations are available, each resulting from the average between eight to twelve human judgments. For a pair of test documents the similarity is computed under the ESA model, using different index collections. The achieved similarities are compared to the human similarity assessments; the correlation is quantified with the Pearson's correlation coefficient. The results achieved by a vector space model based on tf-weighted terms give us a baseline (see Table 1, Row 1).

Experiment 1: Merged Topics. Gabrilovich and Markovitch attribute the success of ESA to the concept hypothesis, which claims that each document of an index collection treats a single concept. By randomly merging 10 Wikipedia articles to a single index document we compiled a topically diffused index collection—without a noteworthy performance deterioration compared to the original Wikipedia index documents (see Table 1, Row 2+3).

Experiment 2: Reuters. The concept hypothesis also claims that an index collection should provide an encyclopedic characteristic. We observe that similar or even higher correlation values are achieved with the Reuters Corpus Volume 1, which definitely does not provide this characteristic (see Table 1, Row 2+4).

Experiment 3: Random Gaussian. Even based on a randomly built index collection where the terms in all documents are $N(0, 1)$ distributed—instead of obeying to Zipf's law or some topical correlation—retrieval results comparable to that of the VSM are achieved. Moreover, since for such a collection no reasonable idf-value is defined, we compare the results also to an ESA model based on the Wikipedia that relies on tf-weighted terms. Note that even these values are nearly achieved (see Table 1, Row 5+6).

[1]This is a small document number for a retrieval experiment, but we resort to this collection for the sake of comparability.

Overall: Index Collection Size. Both the accuracy and the runtime increase with the number of index documents. A reasonable accuracy is achieved with an index collection size $|D_I|$ between 1 000 and 10 000 documents. The runtime is, as expected, linear in $|D_I|$.

The experiments show that both the topical organization and the semantic purity of the index collection is of secondary importance for the retrieval performance: the use of clean Wikipedia articles, merged articles, or Reuters documents is of negligible impact. I.e., the concept hypothesis does not hold. The fact that even a collection of $N(0,1)$ distributed weight vectors does a nearly equally good job in the role of an index collection shows that the α-stability of the term weights may be the actually underlying determinant.[2] However, the size of the index collection matters; it affects both the accuracy and the runtime.

To have a formal basis for the definition of index collection properties and the analysis of the retrieval performance, the next section introduces a formalization of the ESA model. Moreover, this formalization reveals the close connection between the ESA and existing retrieval models.

3. FORMALIZATION OF ESA

Let d be a real-world document, and let $\mathbf{d}$ be a bag-of-word-based representation of d, encoded as a n-dimensional vector of normalized term frequency weights: $||\mathbf{d}|| = 1$. To ensure the comparability between two arbitrary weight vectors $\mathbf{d}_1$ and $\mathbf{d}_2$, their dimensionality as well as their term order is aligned with a universal term vocabulary V that contains all used terms. A set $\mathbf{D}$ of document representations defines a term-document matrix A_D, where each column in A_D corresponds to a vector $\mathbf{d} \in \mathbf{D}$. A_D is an $m \times n$ matrix, i.e., A_D encodes a collection of m documents represented over a vocabulary of size n.

Given a document d we distinguish between its unique base representation $\mathbf{d}$ and a derived collection-relative representation $\mathbf{d}_{|D_I}$. The former is computed solely from the local properties of d, whereas the latter relates d to a particular index collection D_I.

Definition 1 (Collection-Relative Representation) *Let D and D_I be two document collections with representations $\mathbf{D}$, $\mathbf{D}_I$, and term-document matrices A_D and A_{D_I}. Then, the term-document matrix $A_{D|D_I}$ of the collection-relative representation of D with respect to collection D_I is defined as follows:*

$$A_{D|D_I} = A_{D_I}^T \cdot A_D,$$

where A^T designates the matrix transpose of A. D_I is called index collection, A_{D_I} is called translation matrix. Each column in $A_{D|D_I}$ corresponds to the collection-relative representation of a document $d \in D$ and is denoted as $\mathbf{d}_{|D_I}$.

The rationale of this definition becomes clear if one considers that $||\mathbf{d}|| = ||\mathbf{d}'|| = 1$ holds for each weight vector $\mathbf{d} \in \mathbf{D}$ and $\mathbf{d}' \in \mathbf{D}_I$. Hence, each entry in the collection-relative representation $\mathbf{d}_{|D_I}$ of a document $d \in D$ is the cosine similarity between $\mathbf{d}$ and some vector $\mathbf{d}' \in \mathbf{D}_I$. Put another way, d is compared to each document in D_I, and $\mathbf{d}_{|D_I}$ is comprised of the respective cosine similarities.

Between two documents d_1 and d_2, the similarity $\varphi_{ESA}(d_1, d_2)$ under the ESA model is computed as cosine similarity φ of the ESA representations of d_1 and d_2:

$$\varphi_{ESA}(d_1, d_2) := \varphi(\mathbf{d}_{1|D_I}, \mathbf{d}_{2|D_I}) = \varphi(A_{D_I}^T \cdot \mathbf{d}_1, A_{D_I}^T \cdot \mathbf{d}_2).$$

[2]The Gaussian distribution is an example for an α-stable distribution. For details see [8].

This notation of the ESA model is inspired by the generalized vector space model, GVSM [11], which employs knowledge about term co-occurrence within the retrieval process. Under the GVSM the similarity $\varphi_{GVSM}(q, d)$ between a query q and a document d is defined as follows:

$$\varphi_{GVSM}(q, d) = \mathbf{q}^T \cdot G \cdot \mathbf{d},$$

where G is a $n \times n$ term co-occurrence matrix.

The ESA model and the GVSM can be directly transformed into each other when setting $D_I = D$:

$$\begin{aligned} \varphi_{ESA}(q, d) &= \varphi(A_D^T \cdot \mathbf{q}, A_D^T \cdot \mathbf{d}) \\ &= \mathbf{q}^T \cdot A_D \cdot A_D^T \cdot \mathbf{d} \quad = \varphi_{GVSM}(q, d). \end{aligned}$$

The $n \times n$ matrix $A_D \cdot A_D^T$ has a nonzero value in its i-th row and j-th column if there is a document in D where both the i-th and the j-th terms co-occur; hence $A_D \cdot A_D^T = G$.

Analogously, the CL-ESA model [9] and the cross-lingual extension of the GVSM [12] can be transformed into each other.

4. CONCLUSION AND CURRENT WORK

Contrary to accepted belief the ESA concept hypothesis does not hold: the use of clean Wikipedia articles, merged articles, Reuters documents, or even random vectors has a negligible impact on the retrieval performance of the ESA model. However, the size of an index collection matters; it affects both the accuracy and the runtime. Our evaluation provides a guideline for the adjustment of collection-relative models to the needs of a particular retrieval task: put in a nutshell, a reasonable trade-off between accuracy and runtime is achieved with a number of 1 000 - 10 000 index documents.

Our formalization of the ESA model shows its close connection to the generalized VSM. It also opens a perspective to construct index collections for specific information retrieval tasks.

Part of our current work is the identification and quantization of index collection properties that correlate with the observed retrieval performance in practical applications. Based on the expected insights, tailored collections may be constructed for specialized retrieval tasks (e.g., narrow domain versus broad domain) or desired retrieval behavior (e.g., accuracy versus runtime), simply by adjusting mathematical properties of the translation matrix A_{D_I}.

5. REFERENCES

[1] M. Chang, L. Ratinov, D. Roth, and V. Srikumar. Importance of Semantic Representation: Dataless Classification. In *Proc. of AAAI'08*.

[2] O. Egozi, E. Gabrilovich, and S. Markovitch. Concept-Based Feature Generation and Selection for Information Retrieval. In *Proc. of AAAI'08*.

[3] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness of Words and Texts in Wikipedia-derived Semantic Space. Technical Report, Technion, Israel, 2006.

[4] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proc. of IJCAI'07*.

[5] R. Gupta and L. Ratinov. Text Categorization with Knowledge Transfer from Heterogeneous Data Sources. In *Proc. of AAAI'08*.

[6] I. Gurevych, C. Müller, and T. Zesch. What to be? - Electronic Career Guidance Based on Semantic Relatedness. In *Proc. of ACL'07*.

[7] M. Lee, B. Pincombe, and M. Welsh. An Empirical Evaluation of Models of Text Document Similarity. In *Proc. of CogSci'05*.

[8] J. P. Nolan. Stable Distributions—Models for Heavy Tailed Data. http://academic2.american.edu/ jpnolan/stable/stable.html, 2005.

[9] M. Potthast, B. Stein, and M. Anderka. A Wikipedia-Based Multilingual Retrieval Model. In *Proc. of ECIR'08*.

[10] P. Sorg, P. Cimiano. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In *Working Notes for CLEF'08 Workshop*.

[11] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized Vector Space Model in Information Retrieval. In *Proc. of SIGIR'85*.

[12] Y. Yang, J. G. Carbonell, R. D. Brown, and R. E. Frederking. Translingual Information Retrieval: Learning from Bilingual Corpora. *Artificial Intelligence*, 103(1-2):323–345, 1998.