Insights into Explicit Semantic Analysis

Thomas Gottron
University of Koblenz-Landau
56070 Koblenz, Germany
gottron@uni-koblenz.de

Maik Anderka
Bauhaus-Universität Weimar
99421 Weimar, Germany
maik.anderka@uni-weimar.de

Benno Stein
Bauhaus-Universität Weimar
99421 Weimar, Germany
benno.stein@uni-weimar.de

ABSTRACT

Since its debut the *Explicit Semantic Analysis* (ESA) has received much attention in the IR community. ESA has been proven to perform surprisingly well in several tasks and in different contexts. However, given the conceptual motivation for ESA, recent work has observed unexpected behavior. In this paper we look at the foundations of ESA from a theoretical point of view and employ a general probabilistic model for term weights which reveals how ESA actually works. Based on this model we explain some of the phenomena that have been observed in previous work and support our findings with new experiments. Moreover, we provide a theoretical grounding on how the size and the composition of the index collection affect the ESA-based computation of similarity values for texts.

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: Experimentation, Standardization, Theory

1. INTRODUCTION

Gabrilovich and Markovitch introduced the concept of Explicit Semantic Analysis (ESA) in 2007 [4]. The idea underlying ESA is to represent and compare texts (from single terms to entire documents) as vectors in a high dimensional *concept space*. Each dimension in this space corresponds to an explicit semantic concept, where the concepts are considered to be "thematically (or topically) orthogonal". The entries in the concept vector of a given text quantify the associations between the text and the respective concepts. In order to compute such association values, each concept is represented by an index document. Gabrilovich and Markovitch use Wikipedia articles as index documents since Wikipedia covers a wide range of topics, while each article is focused on one topic. This topical focus is needed for the orthogonality property of the concepts and is referred to as *concept hypothesis*.

Since its introduction the ESA model has been adopted quickly in the IR community, which might be attributed to the following facts: the basic idea is pretty simple, Wikipedia is freely available, and the concept representation of documents is easy to interpret. However, recent work reveals that the central concept hypothesis is not required [1]. It has been shown experimentally that instead of the conceptually orthogonal Wikipedia articles, using documents from the Reuters corpus as well as randomly concatenated Wikipedia articles lead to a comparable performance. Even when using random term weights to construct the vectors representing concepts, ESA still achieves quite good performance values. These unexpected results cannot be aligned with the common explanation of ESA and motivate a deeper analysis of how ESA actually works.

Table 1: Correlation of ESA-based similarities with human assessments, depending on index collection, number of index documents, and weighting scheme (reproduced from [1]).

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-idf	0.742	0.784	0.782	0.782	0.781	0.781
Merged topics, tf-idf	0.738	0.767	0.768	0.769	0.769	0.777
Reuters, tf-idf	0.767	0.795	0.802	0.800	0.800	0.800
Random Gaussian, tf	0.703	0.716	0.717	0.717	0.717	0.717

The paper in hand takes a theoretical look at ESA and proposes a probabilistic term weight model to describe the computed similarity values. Our contributions are threefold: (1) we show which information of an index collection is actually exploited by ESA, (2) we investigate the impact of certain features of an index collection (such as size and composition), and (3) we explain the surprisingly good performance of ESA when using index vectors with random term weights. Altogether, the paper gives further evidence that ESA is a variation of the generalized vector space model (GVSM) [8].

2. RELATED WORK

ESA was introduced as an approach to compute the semantic relatedness of terms or short phrases [4]—we explain the technical details in Section 2.2. Meanwhile, ESA has been adopted successfully in many applications. In [3] ESA contributes directly to the estimation of the relevance of documents for a given query. In other settings ESA is used to compute the semantic relatedness of terms [5], which are then used as parameters in other retrieval models (e.g., an extension of BM-25). A cross-lingual extension (CL-ESA) that exploits interlanguage links of Wikipedia articles is covered in [6] and [7]. In [2] it is shown that CL-ESA is superior to other retrieval models which are based on implicit semantics.

Anderka and Stein revisited ESA in [1] and found syntactic parallels to the GVSM—a brief introduction of the GVSM is given in Section 2.1. They also present some initial analysis targeting

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

the impact of the index collection on the performance of ESA, see Table 1. Two findings are quite interesting: First, employing Wikipedia as index collection does not perform best; e.g., using the Reuters corpus results in a better performance, even though the concept hypothesis is not satisfied in this case. Second, using index vectors with random term weights (Random Gaussian) performs nearly as good as using Wikipedia articles. This eventually led the authors to the conclusion, that the initial concept hypothesis in ESA does not hold.

2.1 GVSM in a Nutshell

Under the traditional vector space model (VSM) a document x is represented as a vector $\mathbf{x}$ based on a weighted combination of n-dimensional term vectors $\mathbf{t}_j$:

$$\mathbf{x} = \sum_{j=1}^{m} w(t_j, x) \cdot \mathbf{t}_j \tag{1}$$

where $w(t_j, x)$ is the weight of term t_j in document x. The similarity between two documents x and y is measured via the inner product of the respective vectors $\mathbf{x}$ and $\mathbf{y}$:

$$\begin{aligned} sim_{\text{VSM}}(x, y) &= \langle \mathbf{x},\ \mathbf{y} \rangle && (2) \\ &= \left\langle \sum_{j=1}^{m} w(t_j, x) \cdot \mathbf{t}_j,\ \sum_{k=1}^{m} w(t_k, y) \cdot \mathbf{t}_k \right\rangle && (3) \\ &= \sum_{j=1}^{m} \sum_{k=1}^{m} w(t_j, x) \cdot w(t_k, y) \cdot \langle \mathbf{t}_j, \mathbf{t}_k \rangle && (4) \end{aligned}$$

In the classical VSM, the term vectors $\mathbf{t}_j$ have only a single non-zero entry, are orthogonal and of unit length. This allows to represent documents as term weight vectors and the above product becomes $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^{m} w(t_j, x) \cdot w(t_j, y)$. This simplification also renders the VSM incapable of capturing interdependence of terms.

The generalized vector space model (GVSM) [8] considers term correlations. Based on term co-occurrence information, the values for $\langle \mathbf{t}_j, \mathbf{t}_k \rangle$ are estimated and stored as entries g_{jk} in a matrix G. This allows for maintaining the document representation as weight vectors and formulating the similarity of two documents x and y as follows:

$$sim_{\text{GVSM}}(x, y) = \mathbf{x}^T \cdot G \cdot \mathbf{y} \tag{5}$$

2.2 ESA in a Nutshell

Let D be an index collection of n documents, each of which describing a single concept. Further, let V be a vocabulary with m different terms that occur in D. In the original publication [4] the index collection is build from Wikipedia articles in order to cover a wide range of different concepts. Under the ESA model, a document x is represented as a concept vector $\mathbf{u}$, where each entry in $\mathbf{u}$ corresponds to the inner product of $\mathbf{x}$ and $\mathbf{d}_i$, with $d_i \in D$:

$$\mathbf{u}^T = \left(\langle \mathbf{d}_1, \mathbf{x} \rangle, \langle \mathbf{d}_2, \mathbf{x} \rangle, \ldots, \langle \mathbf{d}_n, \mathbf{x} \rangle \right), \tag{6}$$

where $\mathbf{x}$ and $\mathbf{d}_i$ are the VSM representations of x and d_i, which are m-dimensional vectors containing a weight for each term in V. Here, the original publication suggests to use a standard tf-idf weighing scheme. The inner product $\langle \mathbf{d}_i, \mathbf{x} \rangle$ corresponds to the similarity of x and d_i under the VSM; it quantifies the association between x and the concept which is defined by d_i.

Under the ESA model, the similarity between two documents x and y is defined by the cosine measure that is applied to the respective concept vectors $\mathbf{u}$ and $\mathbf{v}$:

$$sim_{\text{ESA}}(x, y) = \cos(\mathbf{u}, \mathbf{v}) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{|\mathbf{u}| \cdot |\mathbf{v}|} \tag{7}$$

3. THEORETICAL ANALYSIS OF ESA

The findings in [1] motivate a deeper analysis of ESA and the need to understand what actually happens when comparing vectors in the concept space. We start with a reformulation of sim_{ESA} (see Equation 7) using the definitions of the concept vector entries:

$$\begin{aligned} sim_{\text{ESA}}(x, y) &= \frac{1}{|\mathbf{u}| \cdot |\mathbf{v}|} \sum_{i=1}^{n} \langle \mathbf{d}_i, \mathbf{x} \rangle \cdot \langle \mathbf{d}_i, \mathbf{y} \rangle && (8) \\ &= \frac{1}{|\mathbf{u}| \cdot |\mathbf{v}|} \sum_{j=1}^{m} \sum_{k=1}^{m} w(t_j, x) \cdot w(t_k, y) \cdot g_{jk} && (9) \end{aligned}$$

where $g_{jk} = \sum_{i=1}^{n} w(t_j, d_i) \cdot w(t_k, d_i)$. Essentially this leads to a structural reformulation of ESA as GVSM, which is also observed in [1]. The g_{jk} are of particular interest as they contribute the entries in matrix G, see Equation 5. In Section 3.1 we will take a closer look at how these values evolve.

First we note that the g_{jk} also appear in the length normalization:

$$\begin{aligned} |\mathbf{u}| &= \sqrt{\sum_{i=1}^{n} \left(\langle \mathbf{d}_i, \mathbf{x} \rangle \right)^2} && (10) \\ &= \sqrt{\sum_{i=1}^{n} \left[\sum_{j=1}^{m} w(t_j, d_i) \cdot w(t_j, x) \right]^2} && (11) \\ &= \sqrt{\sum_{j=1}^{m} \sum_{k=1}^{m} w(t_j, x) \cdot w(t_k, x) \cdot g_{jk}} && (12) \end{aligned}$$

The recurrence of the g_{jk} is of interest because it allows us to scale G with a factor $a > 0$ without affecting the similarity measure. This can be seen by extending each g_{jk} into $a \cdot \frac{1}{a} \cdot g_{jk}$ and factoring $\frac{1}{a}$ out of the sum:

$$sim_{\text{ESA}}(x, y) = \underbrace{\frac{\frac{1}{a}}{|\mathbf{u}| \cdot |\mathbf{v}|}}_{*} \cdot \sum_{j=1}^{m} \sum_{k=1}^{m} w(t_j, x) \cdot w(t_k, y) \cdot (a \cdot g_{jk})$$

The same can be done in the calculations of the vector lengths for $\mathbf{u}$ and $\mathbf{v}$ in the denominator of the fraction marked with the asterisk ($*$). The resulting $\sqrt{\frac{1}{a}} \cdot \sqrt{\frac{1}{a}}$ can be canceled against the $\frac{1}{a}$ in the numerator. I.e., we can scale the g_{jk} without changing the values of the similarity measure. We will use a scale of $\frac{1}{n}$ to counterbalance the increasing values in the entries of the matrix for growing n. Altogether, we will work with a scaled version G' of the matrix with the entries:

$$g'_{jk} = \frac{1}{n} \sum_{i=1}^{n} w(t_j, d_i) \cdot w(t_k, d_i) \tag{13}$$

Based on G', the ESA similarity measure can be formulated in analogy to the GVSM:

$$sim_{\text{ESA}}(x, y) = \frac{1}{|\mathbf{u}| \cdot |\mathbf{v}|} \cdot \mathbf{x}^T \cdot G' \cdot \mathbf{y}, \tag{14}$$

where the only difference to Equation 5 is the length normalization factor;[1] $|\mathbf{u}|$ can be represented in the form $\sqrt{\mathbf{x}^T \cdot G' \cdot \mathbf{x}}$.

[1] In the original GVSM this factor is omitted since the weight vectors are considered to be normalized.

3.1 Probabilistic Model of Term Weights

In order to derive conclusions about the values of the g'_{jk}, we model the term weights $w(t_j, d_i)$ for the documents $d_i \in D$ as random variables. Typically, term weights are based on features such as term frequency, document frequency, or document length. In many (probabilistic) IR models, these features are modeled as random variables themselves. So, the idea to assume the term weights to be random variables is not far fetched.

The representation of a document d_i is assumed to be a tuple of random variables $(W_{1i}, W_{2i}, \ldots, W_{mi})$, where W_{ji} models the weight of term t_j. For each term t_j we consider all W_{ji} to be i.i.d. over the documents. This implies two assumptions: (1) the term weights in a particular document do not depend on weights in other documents, and (2) the joint distribution of all weights is the same for each document.

Both assumptions can be derived directly from our document model. We consider a document as the realization of the terms occurring in it; i.e., terms form a document—and not the other way around. Of course, there will be latent influences in a document (author, topic, intended auditorium, language style, etc.) which influence the choice and joint distribution of the words. These can be captured by the interdependence of the W_{ji} within a document, about which we explicitly make no assumption. Based on this document model, the g'_{ik} represent the mean over products of random variables W_{ji}:

$$g'_{jk} = \frac{1}{n} \sum_{i=1}^{n} W_{ji} \cdot W_{ki} \tag{15}$$

Since the W_{ji} are i.i.d. with respect to the documents, the law of large numbers provides the almost sure convergence for $n \to \infty$:

$$g'_{jk} \xrightarrow{a.s.} E(W_{ji} \cdot W_{ki}) = Cov(W_{ji}, W_{ki}) + E(W_{ji}) \cdot E(W_{ki}) \tag{16}$$

I.e., in general the entries g'_{jk} converge to a representation that captures the covariance of the term weights. This suits the concept of the matrix G in the GVSM, where the entries are supposed to represent the correlation of term vectors; however, the approach to determine the correlation values is different.[2] A second interesting observation is that a single g'_{jk} determines the measure of semantic relatedness for the terms t_j and t_k. So, ESA considers the common distribution of the weights of those two terms in the index collection to measure semantic relatedness. The covariance denotes the deviation from a random and uncorrelated co-occurrence of terms. Hence, rather than exploiting orthogonal concepts, ESA actually benefits from redundant usage of terms in documents. This reveals that there is no substantive evidence for the benefit of the concept hypothesis.

4. EXPERIMENTS

The ESA formulation of the previous section will help us to investigate properties of the index collection. In particular, we investigate the document number and the topic composition, and we explain the fact that ESA still works with random weight vectors. Beyond explaining observations from previous work, we extend the results reported in [1] to empirically support our theoretical findings.

4.1 Index Collection Size

From a practical point of view the index collection size affects the runtime of ESA, since a given document needs to be compared against all index documents. Given this constraint, a small index collection is favorable. However, given the convergence in Equation 16, a large index collection is favorable since it reduces the variance of the g'_{jk} entries. A small variance is desired because it eliminates random fluctuations, which might be caused by outlier documents in the index collection.

This hypothesis is supported by the experiment results shown in Table 1: With increasing number of index documents the performance of ESA becomes more stable. Only if randomly concatenated Wikipedia articles are used as index documents (merged topics), the performance shows a significant increase at $200,000$ documents. Within the merged topics collection each index documents is generated by concatenating 10 randomly drawn Wikipedia articles with at least 1,000 words. However, given the size of the index collection several articles must have been reused; i.e., the assumption of the W_{ji} to be independent across the documents cannot hold, which might explain this artifact in the data.[3]

4.2 Topic Composition

Equation 16 and the reformulation of ESA as a special case of the GVSM in Equation 14 show that the index collection essentially provides term correlation information. Moreover, ESA operates on the overlap of the vocabulary defined by the documents x and y, as well as the vocabulary defined by the index documents. The summands in Equation 9 contain the term $w(t_j, x) \cdot w(t_k, y) \cdot g_{jk}$, which contributes to the similarity score if and only if $w(t_j, x)$, $w(t_k, y)$, and g_{jk} are nonzero. I.e., one will benefit from the correlation information of two terms only if they occur in different documents x, y, as well as in the index collection. One hence might argue that an index collection that covers the same topic domain like the compared documents will be more beneficial than a collection about a different topic, just because of the vocabulary overlap. Again, this hypothesis is supported by the experimental results in Table 1. Note that using the Reuters corpus as index collection entails a better performance than using Wikipedia articles. A similar observation was also made by Müller and Gurevich in [5]. The results in Table 1 are based on a test corpus of news documents from the Australian Broadcasting Corporation's news mail service, which focus on international politics; see [1] for further details.

It is important to note that the wide topic range in Wikipedia can also be a disadvantage: even if the vocabulary overlap is large, homonyms from different topics will introduce noise and distortions in the correlation information. For instance, the term "capital" has several meanings (capital city, financial capital, capital letters or top part of a column). At a semantic level the meanings of this term are individually correlated with different other terms. At a syntactical level the semantic differences are lost and the correlations of all meanings with, for instance, the term "government", get blurred. In a narrow domain corpus such as Reuters it is likely that the number of different meanings of a term is low. By contrast, Wikipedia contains most—if not all—different meanings and thus introduces much more noise in the correlation weights of G'. I.e., the original motivation of using Wikipedia because of its topic range can be counterproductive when addressing a particular topic domain.

[2] The details of how the entries in the matrix G are calculated in the original GVSM approach go beyond the scope of this paper and can be found in [8].

[3] In January 2010 Wikipedia contained less than 2,000,000 articles with 1,000 words. http://stats.wikimedia.org/EN/TablesWikipediaEN.htm

Table 2: Correlation of ESA-based similarities with human similarity assessments, depending on index collection and number of index documents. As weighing scheme tf-idf is used.

Index collection	**Number of index documents**			
	1,000	10,000	50,000	100,000
VSM (baseline)	0.717	0.717	0.717	0.717
DMOZ - sport	0.700	0.748	0.752	0.752
DMOZ - science	0.703	0.736	0.743	0.743
DMOZ - news	0.754	0.784	0.788	-
Wikipedia - arts	0.709	0.757	0.760	0.759
Wikipedia - science	0.726	0.776	0.780	0.779
Wikipedia - politics	0.776	0.798	0.798	0.798

To substantiate this hypothesis we conduct a focused crawl of documents from both the open directory project (DMOZ) and Wikipedia. We create different index collections of up to 100,000 documents with a topical focus on sports, science and news for DMOZ, as well as on arts, science and politics for Wikipedia.[4] For each index collection we apply the same evaluation procedure as in [1], using the test corpus of news documents from the Australian Broadcasting Corporation's news mail service. The results are shown in Table 2. The best results are achieved with index collections based on news documents and politics documents, which shows that ESA benefits from an index collection about the same topic domain like the documents that are to be compared.

4.3 Random Weight Vectors

The authors of [1] report that an index collection with term weights created by a random process yield good results as well. The term weights were chosen to be independently $N(0,1)$ distributed. We can directly use these parameter settings here in order to obtain $\mathit{Var}(W_j) = 1$, $E(W_j) = 0$ and $\mathit{Cov}(W_j, W_j) = 0$ for $j \neq k$. Using this in the Equation 16 results for $j \neq k$ in:

$$g'_{jk} \xrightarrow{a.s.} \mathit{Cov}(W_{ji}, W_{ki}) + E(W_{ji}) \cdot E(W_{ki}) = 0 \quad (17)$$

The elements g'_{jj} on the main diagonal are treated differently:

$$g'_{jj} \xrightarrow{a.s.} \mathit{Cov}(W_{ji}, W_{ji}) + E(W_{ji}) \cdot E(W_{ji}) = 1 \quad (18)$$

For $n \to \infty$ we observe that G' becomes a $m \times m$ unit matrix I. A unit matrix in the GVSM corresponds to the classical VSM with orthonormal term vectors. Also this theoretical finding is reflected by the results shown in Table 1. With increasing size of the randomly generated index collection (Random Gaussian) the performance converges towards a value of 0.717, which is the same value one obtains with the classical VSM baseline.

5. CONCLUSIONS

As main contribution we develop a view of ESA that shows its conceptual equivalence with the GVSM, independent of the choice of term weights for the index documents. On the basis of this reformulation and by employing a probabilistic model for term weights, we show that ESA essentially captures term correlation information from the index collection. This correlation information is exploited to match different terms in compared documents, which also shows that the concept hypothesis has no (obvious) mathematical basis.

In addition, we develop a deeper understanding of the influence of the index collection: (1) Increasing the number of index documents causes the correlation values for terms to converge, and hence, larger index collections provide more reliable correlation information. (2) The application of ESA in a specific domain benefits from taking an index collection from the same topic domain while, on the other hand, a "general topic corpus" such as Wikipedia introduces noise. Finally, our model is able to explain the behavior of ESA when using random term weights in index documents: with increasing document number the ESA-based similarities converge to the classical VSM using the cosine measure.

Our theoretical findings are supported by empirical results presented in related work. We also support our findings by new empirical analysis using different topically focused index collections. The paper also explains the behavior of ESA that seemed to be contradictory to the assumptions previously made on how ESA works.

Given these insights into ESA, future work will deal with the analysis and development of measures that can support the choice of a suitable index collection for a given task and context. Vocabulary overlap or a comparison of the distribution of term weights are first candidates to be considered. A second step is the composition within the index collection. Given the impact on runtime performance, it is of interest to reduce the size of the index collection while maintaining the effectiveness. Finally, also the impact of the findings presented here on CL-ESA should be explored.

6. REFERENCES

[1] M. Anderka and B. Stein. The ESA retrieval model revisited. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 09)*, pages 670–671. ACM, 2009.

[2] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 09)*, pages 1513–1518. Morgan Kaufmann Publishers Inc., 2009.

[3] O. Egozi, E. Gabrilovich, and S. Markovitch. Concept-based feature generation and selection for information retrieval. In *Proceedings of the 23rd National Conference on Artificial Intelligence (AAAI 08)*, pages 1132–1137. AAAI Press, 2008.

[4] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 07)*, pages 1606–1611. Morgan Kaufmann Publishers Inc., 2007.

[5] C. Müller and I. Gurevych. Semantically enhanced term frequency. In *Advances in Information Retrieval: Proceedings of the 32nd European Conference on IR Research (ECIR 10)*, volume 5993 of *Lecture Notes in Computer Science*, pages 598–601. Springer, 2010.

[6] M. Potthast, B. Stein, and M. Anderka. A Wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval: Proceedings of the 30th European Conference on IR Research (ECIR 08)*, volume 4956 of *Lecture Notes in Computer Science*, pages 522–530. Springer, 2008.

[7] P. Sorg and P. Cimiano. Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes for the CLEF 2008 Workshop (CLEF 08)*, 2008.

[8] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized vector spaces model in information retrieval. In *Proceedings of the 8th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 85)*, pages 18–25. ACM, 1985.

[4] The DMOZ category *news* comprises only 67,809 documents.