Tjoa, Wagner (Eds.): Proceedings of the Dexa '09, Linz
20th International Conference on Database and Expert Systems Applications
ISBN 978-0-7695-3763-4, pp. 383-387, ©IEEE 2009.

Collection-Relative Representations
A Unifying View to Retrieval Models

Benno Stein and Maik Anderka
Faculty of Media, Media Systems
Bauhaus-University Weimar
99421 Weimar, Germany
Email: <first>.<last>@uni-weimar.de

***Abstract*—Various retrieval models have been developed and analyzed so far, but less research aims to an integration of the different models within a common framework. This paper introduces the idea of collection-relative retrieval models, a paradigm where several important retrieval models fit in. Our unifying view helps to better understand retrieval models, and it can be considered as a step towards a common theoretical framework for text retrieval.**

Collection-relative retrieval models employ a so-called index collection. We present an evaluation that shows how particular characteristics of the underlying index collection affect the retrieval performance of a collection-relative model. Based on such insights tailored index collections can be constructed in order to address specialized retrieval tasks.

I. Introduction

Retrieval models can be considered as heuristics that operationalize the *probability ranking principle* [8]: "Given a query q, the ranking of documents according to their probabilities of being relevant to q leads to the optimum retrieval performance." Different retrieval models rely on different paradigms to assess relevance: we find, among others, models that quantify document similarity, models that project the term space into a concept space, models that quantify term generation likelihoods, or models that exploit human relevance information. Note that this distinction is independent of the probability ranking principle—to which virtually all models obey.[1]

Less research aims to an integration of different retrieval models within a common framework. One of the exceptions is [7], where the authors propose a language modeling approach to integrate "models of document indexing" and "[probabilistic] models of document retrieval". Similarly, Crestani [1] provide a comparative view onto the class of probabilistic retrieval models. The model comparison of Srikanth [12] is insightful but not intended as a formalization. In [9] a "general matrix framework for modelling information retrieval" is proposed, but the authors do not show how existing retrieval models can be interpreted within the framework.

[1]The principle cannot be applied to all kinds of retrieval tasks. In comment ranking, for example, the differential information gain must be considered.

In this paper we develop the idea of collection-relative retrieval models, a paradigm where several existing retrieval models fit in. A collection-relative representation of a document d results from the comparison of d relative to an entire document collection, the *index collection*. Table I lists a number of well-known retrieval models, from which a large part can be interpreted as being collection-relative at heart.

Our framework for collection-relative retrieval models is a generalization of the explicit semantic analysis (ESA) model of Gabrilovich and Markovitch [3], [2]. The ESA representation of a document d is understood as a projection of d into the concept space spanned by a foreign document collection D_I, which is called index collection here. The supposed rationale of the ESA model is that each document in D_I functions as a concept to which the original document d is compared. In [3] both Wikipedia and the Open Directory Project are used in the role of D_I.

The retrieval performance of a collection-relative retrieval model can be controlled by the index collection, e.g., by its size, domain, or topical organization. With deeper knowledge about such relations tailored collections can be constructed for specialized retrieval tasks (e.g., narrow domain versus broad domain) or desired retrieval behavior (e.g., accuracy versus runtime).

A. Contributions

Main contribution of the paper in hand is the idea of collection-relative retrieval models, which is formally introduced in Section II. Moreover, we interpret several classical retrieval models as special instances of collection-relative models. Section III presents an evaluation that is intended to serve as a guideline for the adjustment of collection-relative models to the needs of a given retrieval task. In particular we show that, contrary to common belief, both the topical organization and the semantic purity of the index collection are of secondary importance for the retrieval performance. However, the size of an index collection matters; it affects the accuracy and the runtime of collection-relative retrieval.

II. Collection-Relative Retrieval

Let d be a real-world document, and let $\mathbf{d}$ be a bag-of-word-based representation of d encoded as 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 $n \times m$ 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 the derived collection-relative representations $\mathbf{d}_{|D_1}, \ldots, \mathbf{d}_{|D_k}$. The former is computed solely from the local properties of d, whereas each of the latter representations relates d to a particular index collection $D_I \in \{D_1, \ldots, D_k\}$.

Definition 1 (Collection-Relative Representation) *Let D and D_I be two document collections with representations $\mathbf{D}$ and $\mathbf{D}_I$, and with 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 the weight vectors $\mathbf{d} \in \mathbf{D}$ and $\mathbf{d}' \in \mathbf{D}_I$. Hence, each entry in the collection-relative representation $\mathbf{d}_{|D_I}$ of a document $d \in D$ corresponds to the cosine similarity between $\mathbf{d}$ and some vector $\mathbf{d}' \in \mathbf{D}_I$. Put another way, d is compared to all documents 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 φ_{CRRM} under the collection-relative retrieval model, CRRM, is computed as cosine similarity φ of the collection-relative representations of d_1 and d_2:

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

When given a query q and a collection D from which the most similar document d^* wrt. q is desired, collection-relative retrieval is straightforward:

$$d^* = \underset{d \in D}{\operatorname{argmax}} \ \varphi_{CRRM}(q, d) \tag{1}$$

A. Collection-Relative Interpretation of Retrieval Models

Several classical retrieval models can be interpreted as special instances of a collection-relative retrieval model, see Table I. This fact will be illustrated in the following with the conventional and the generalized vector space model [11],

Table I: Classification of known retrieval models.

	Retrieval model	Document representation	Foundation	Modeling focus
collection-relative retrieval models	vector space model, VSM	*tf*-weighted terms, *tf·idf*-weighted terms	empirical	similarity: $\varphi(\mathbf{q}_{\|D_{TF}}, \mathbf{d}_{\|D_{TF}})$
	generalized VSM, GVSM	query expansion based on a term correlation matrix	empirical	similarity: $\varphi(\mathbf{q}_{\|D}, \mathbf{d}_{\|D})$
	latent semantic indexing, LSI	weighted concepts based on a singular value decomposition, SVD	empirical	similarity: $\varphi(\mathbf{q}_{\|D_{\overline{LSI}}}, \mathbf{d}_{\|D_{LSI}})$
	explicit semantic analysis, ESA	weighted concepts based on the similarities to an index collection D_{ESA}	empirical	similarity: $\varphi(\mathbf{d}_{1\|D_{ESA}}, \mathbf{d}_{2\|D_{ESA}})$
	cross-language ESA, CL-ESA	like ESA, but uses on two aligned index collections D_{ESA_1} and D_{ESA_2}	empirical	similarity: $\varphi(\mathbf{d}_{1\|D_{ESA_1}}, \mathbf{d}_{2\|D_{ESA_2}})$
	folding-in LSI	like LSI, but the document d was not used within in the SVD	empirical	similarity: $\varphi(\mathbf{q}_{\|D_{\overline{LSI}}}, \mathbf{d}_{\|D_{\overline{LSI}}})$
probabilistic models	binary independence, BIR	weighted terms from a maximum likelihood estimation for relevance	statistical	relevance: $\rho(R = 1 \mid \mathbf{q}, \mathbf{d})$
	2-Poisson	weighted terms drawn from a document-specific poisson mixture	statistical	relevance: $\rho(R = 1 \mid \mathbf{q}, \mathbf{d})$
generative (language) models	unigram language model	weighted terms that model the likelihood of a query generation	empirical	generation likelihood: $L(\mathbf{q} \mid \mathbf{d})$
	latent dirichlet allocation, LDA	topic-driven generation of weighted terms using a dirichlet distribution	statistical	generation likelihood: $L(\mathbf{d}_1 \mid \mathbf{d}_2)$

[13], the LSI model [4], the ESA model [3], and the CL-ESA model [10]. An important distinction is whether a foreign collection or the retrieval collection itself, i.e., the collection against the query q is matched is used as basis for the translation matrix A_{D_I} (see Figure 1).

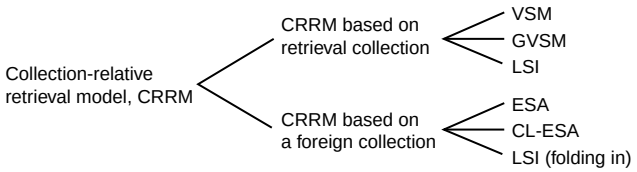

Figure 1: Taxonomy of collection-relative retrieval models.

Vector Space Model, VSM. $A_{D_I} := A_{D_{TF}}$, where D_{TF} is an index collection consisting of $|V|$ one-word documents:

$$\forall\ d_i \in D_{TF} : d_i = \{t_i\},\ \text{with } 1 \leq i \leq |V|,\ t_i \in V$$

Rationale: the translation matrix $A_{D_{TF}}$ is comprised of all n-dimensional unit vectors and forms a permutation matrix. The term-document matrix of a collection D represented under the VSM is given as

$$A_{D|D_{TF}} = A^T_{D_{TF}} \cdot A_D = A_D$$

Analogously, $A_{D_{IDF}}$ denotes the translation matrix of a collection D under the VSM that uses $tf{\cdot}idf$-weighted terms.

Generalized Vector Space Model, GVSM. $A_{D_I} := A_D$. Rationale: the generalized vector space model expands a query q using the term co-occurrence matrix $A_D \cdot A^T_D$, where D is the retrieval collection. Retrieval under the GVSM hence means to evaluate $\mathbf{q}^T \cdot A_D \cdot A^T_D \cdot \mathbf{d}$, which can be rewritten as $A^T_D \cdot \mathbf{q} \cdot A^T_D \cdot \mathbf{d}$. The term-document matrix of a collection D represented under the GVSM is given as

$$A_{D|D} = A^T_D \cdot A_D$$

Latent Semantic Indexing, LSI. $A_{D_I} := A_{D_{LSI}} = \Sigma^{-1}_{D_k} \cdot U^T_{D_k}$, where Σ_{D_k} is a diagonal matrix with the k largest singular values from A_D, while U_{D_k} contains the corresponding singular vectors. The term-document matrix of D represented under the LSI retrieval model is given as

$$A_{D|D_{LSI}} = A^T_{D_{LSI}} \cdot A_D = \Sigma^{-1}_{D_k} \cdot U^T_{D_k} \cdot A_D$$

Explicit Semantic Analysis, ESA. $A_{D_I} := A_{D_{ESA}}$, where D_{ESA} is a subset of Wikipedia used as index collection, encoded as normalized $tf \cdot idf$ vectors. The term-document matrix of D under the ESA retrieval model is given as

$$A_{D|D_{ESA}} = A^T_{D_{ESA}} \cdot A^T_{D_{IDF}} \cdot A_D$$

Cross-Language Explicit Semantic Analysis, CL-ESA. In cross-language information retrieval one is given a query q in language L_1 and a document collection D in language L_2, where the task is to find the most similar document d^* in D wrt. q.

Let, for example, D_{L_1} and D_{L_2} designate two sets of Wikipedia articles in the languages L_1 and L_2, aligned using the Wikipedia inter-language links and encoded as normalized $tf{\cdot}idf$ vectors. When representing q and D under the CL-ESA model, $A_{D_{L_1}}$ and $A_{D_{L_2}}$ are in the role of A_{D_I}. I.e., the CL-ESA representation of q is given as

$$\mathbf{q}_{|D_{L_1}} = A^T_{D_{L_1}} \cdot A^T_{D_{IDF}} \cdot \mathbf{q}$$

Likewise, the term-document matrix of D under the CL-ESA model is given as

$$A_{D|D_{L_2}} = A^T_{D_{L_2}} \cdot A^T_{D_{IDF}} \cdot A_D$$

Due to the alignment of D_{L_1} and D_{L_2} the CL-ESA representations of q and D are comparable. In analogy to Equation 1 the task of cross-language information retrieval can hence be formulated as follows:

$$d^* = \underset{d \in D}{\operatorname{argmax}}\ \varphi(\mathbf{q}_{|D_{L_1}}, \mathbf{d}_{|D_{L_2}})$$

III. Empirical Evaluation

Gabrilovich and Markovitch [3] 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.

A. Experiments and Results

We use the same experimental set-up as the authors in [3], a test collection consisting of 50 documents from the Australian Broadcasting Corporation's news mail service [5]. 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, based on different index collections. As index collection we use a snapshot of the English Wikipedia from September 2008, the Reuters Corpus Volume 1, and a set of random Gaussian vectors. The computed similarities are compared to the human similarity assessments; the correlation is quantified with Pearson's correlation coefficient.

This experimental set-up and the small document number is unusual in information retrieval, but we resort to this collection for the sake of comparability. In addition, we repeat these experiments with the TREC-8 test collection [14]. This collection contains 528 155 documents and 50 queries for which human relevance assessments are available. For each query the similarities to all documents are computed under

Table II: 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

Table III: The MAP achieved on TREC-8 with ESA based on different index collections depending on the number of index documents.

Index collection	number of index documents	
	1 000	10 000
VSM (baseline)	0.110	0.110
Wikipedia, $tf \cdot idf$	0.124	0.160
Merged Topics, $tf \cdot idf$	0.120	0.168
Reuters, $tf \cdot idf$	0.138	0.164
Wikipedia, tf	0.111	0.141
Random Gaussian, tf	0.109	0.132

the ESA model, taking different index collections. The result is a list of documents, ranked according to their similarities to the query. The performance is measured by Mean Average Precision, MAP. Note that no further techniques are applied to improve the ranking, e.g., query expansion or relevance feedback: objective is not to produce a perfect ranking, but to compare the effect of the different index collections underlying ESA.

For both test collections the ranking results that were achieved with the vector space model based on tf-weighted terms define the baseline; see Table II and Table III, 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 into a single index document we compile a topically diffused index collection—without observing a noteworthy performance deterioration compared to the original Wikipedia index documents (see Table II and Table III, 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 II and Table III, Row 2+4).

Experiment 3: Random Gaussian. Even under a set of $N(0,1)$ distributed vectors—without obeying 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 still these values are nearly achieved (see Table II and Table III, Row 5+6).

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 $|D_I|$.

IV. Conclusion and Current Work

A large part of well-known retrieval models can be interpreted as being collection-relative, a paradigm which is introduced in this paper. The collection-relative representation of a document d results from the comparison of d relative to an index collection D_I.

Our experiments with both the original ESA test collection and the TREC-8 collection show that the topical organization as well as the semantic purity of an index collection are of secondary importance for the retrieval performance: clean Wikipedia articles, merged articles, Reuters documents, or even random vectors lead to similar results. The fact that a collection of $N(0,1)$ distributed weight vectors does an 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] Altogether, we conclude that the concept hypothesis does not hold.

Our evaluation provides a guideline for the adjustment of collection-relative models to the needs of a particular retrieval task. If, for example, a high retrieval quality is desired, $|D_I|$ should be 50 000 to 100 000; if a high retrieval efficiency is desired, $|D_I|$ should be 1 000. At a lower number of index documents the retrieval quality deteriorates significantly.

Part of our current work is the identification and quantization of characteristics of the translation matrix, A_{D_I}, which correlate with the observed retrieval performance in practical applications. Based on such insights tailored index collections shall be constructed for specialized retrieval tasks or desired retrieval behavior, simply by optimizing certain mathematical matrix properties of A_{D_I}.

References

[1] F. Crestani, M. Lalmas, C. van Rijsbergen, and I. Campbell. "Is this Document Relevant?. . .Probably": A Survey of Probabilistic Models in Information Retrieval. In *ACM Comput. Surv.*, 30(4):528–552, 1998.

[2] E. Gabrilovich. *Feature Generation for Textual Information Retrieval Using World Knowledge*. PhD thesis, Technion, Israel, 2006.

[2]The Gaussian distribution is an example for an α-stable distribution. For details see [6].

[3] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proc. of IJCAI 2007*.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by Latent Semantic Analysis. In *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[5] M. Lee, B. Pincombe, and M. Welsh. An Empirical Evaluation of Models of Text Document Similarity. In *Proc. of CogSci 2005*.

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

[7] J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *Proc. of SIGIR 1998*.

[8] S. E. Robertson. *The Probability Ranking Principle in IR*. Morgan Kaufmann Publishers Inc., San Francisco, 1997.

[9] T. Rölleke, T. Tsikrika, and G. Kazai. A General Matrix Framework for Modelling Information Retrieval. In *Information Processing & Management*, 42(1):4–30, 2006.

[10] M. Potthast, B. Stein, and M. Anderka. A Wikipedia-Based Multilingual Retrieval Model. In *Proc. of ECIR 2008*.

[11] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. In *Commun. ACM*, 18(11):613–620, 1975.

[12] M. Srikanth. *Exploiting Query Features in Language Modeling Approach for Information Retrieval*. PhD thesis, State University of New York at Buffalo, USA, 2004.

[13] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized vector spaces model in information retrieval. In *Proc. of SIGIR 1985*.

[14] E. M. Voorhees and D. K.-Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). In *Proc. of TREC-8*, 1999.