Pseudo Descriptions for Meta-Data Retrieval

Tim Gollub
Bauhaus-Universität Weimar
Weimar, Germany
tim.gollub@uni-weimar.de

Erdan Genc
Bauhaus-Universität Weimar
Weimar, Germany
erdan.genc@uni-weimar.de

Nedim Lipka
Adobe Systems
San Jose, USA
lipka@adobe.com

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

ABSTRACT

Search in meta-data is challenging due to the sparsity of the available textual information. To alleviate the sparsity problem, the paper in hand evolves from the existing document expansion paradigm and proposes pseudo-descriptions as a new paradigm. Instead of encoding paradigmatic term relations implicitly in an expansion vector, we generate an explicit cohesive text field for meta-data records that describes the entity associated with the record. In contrast to document expansions, pseudo-descriptions allow to reveal why a certain document is considered relevant although the original meta-data does not contain the query terms. Moreover, they are easier to operationalize and facilitate the use of sophisticated retrieval features such as phrase search and query term proximity. To generate pseudo-descriptions, we propose a relevance dependent strategy that depends on the search engine result pages obtained from issuing the meta-data as a search query to a designated reference search engine. To demonstrate the validity of the pseudo-description paradigm, we experiment with different TREC collections where we withhold the content information to simulate a meta-data retrieval scenario. Though retrieval with full content information remains superior, our approach achieves retrieval performance improvements en par with document expansion.

ACM Reference Format:
Tim Gollub, Erdan Genc, Nedim Lipka, and Benno Stein. 2018. Pseudo Descriptions for Meta-Data Retrieval. In *2018 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '18), September 14–17, 2018, Tianjin, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3234944.3234957

1 INTRODUCTION

Search in meta-data collections is a common information retrieval scenario. The scenario is given whenever the entities of a collection are not linguistic but designate products, persons, or multimedia, or if the entities are linguistic but not available to the search engine provider. Prominent providers of meta-data search engines are e-commerce sites indexing product information, or libraries/archives indexing bibliographic records. From a search engine developer's point of view, meta-data collections are pleasant to work with in comparison to "standard" collections such as the Web: They usually can be obtained easily, are free of charge, professionally maintained, provided in a single standardized format, and compact, rendering both indexing and query processing efficient.

1.1 Retrieval Challenges

The compactness of meta-data enables efficient search, but negatively impacts retrieval effectiveness. Due to the sparsity of lexical clues, meta-data retrieval often suffers noticeably from the so called vocabulary mismatch problem and the headroom problem [4].

The vocabulary mismatch problem arises since standard retrieval models ignore paradigmatic correlations between terms. Terms are treated as orthogonal basis vectors of an n-dimensional vector space, in which documents d and queries q are represented as vectors $\mathbf{q}$ and $\mathbf{d}$. The relevance score $\mathrm{RSV}(q, d)$ is commonly computed by a multiplication of the two vectors, and hence zero if no terms match:

$$\mathrm{RSV}(q, d) = \mathbf{q}^\intercal \mathbf{d} \quad with \quad \mathbf{q}, \mathbf{d} \in \mathbb{R}^{n \times 1}$$

The less lexical information is provided in a meta-data collection, the more likely vocabulary mismatches will harm retrieval effectiveness.

The headroom problem arises since standard retrieval models rely on relative term frequency information to determine the result rank of a document. In collections with sparse lexical clues such as meta-data records, the term frequencies are likely to have Boolean character, i.e., terms occur either not or once. As a consequence, small variations in the document length determine the relevance ranking, which is, as pointed out by Efron et al. [4], often not sensible. Figure 1 illustrates how limited lexical information harms the search effectiveness in the collections used in our experiments.

1.2 Expansion Approaches

To tackle the problems that arise from sparse lexical information, query and document expansion approaches have been proposed. As detailed in Section 2, based on computed paradigmatic term relationships, these approaches add terms to the initial vector representations $\mathbf{q}$ or $\mathbf{d}$ to alleviate the vocabulary mismatch problem, and apply term weight smoothing to alleviate the headroom problem. The computed expansion vectors $\mathbf{q}'$ or $\mathbf{d}'$ are then factored into the

ICTIR '18, September 14–17, 2018, Tianjin, China

ACM ISBN 978-1-4503-5656-5/18/09…$15.00
https://doi.org/10.1145/3234944.3234957

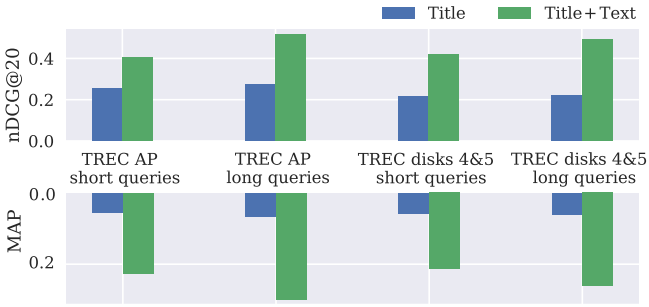

Figure 1: Retrieval effectiveness of Elastic Search with and without content information for two news collections. The nDCG@20 and MAP drop significantly if no content information is available.

relevance function $\mathrm{RSV}(q, d)$, typically using a weighted sum. E.g., the relevance scoring for a document expansion then reads:

$$\mathrm{RSV}_{\mathbf{d}'}(q, d) = \mathbf{q}^\intercal (w_1\, \mathbf{d} + w_2\, \mathbf{d}') \tag{1}$$

1.3 The Pseudo Description Paradigm

Common to all document expansion approaches is the fact that the lexical information that gives rise to the paradigmatic relations encoded in the expansion vectors is not retained explicitly. However, this lexical information may be valuable to have for a meta-data search engine: It would allow to provide search result snippets that reveal why a document is considered relevant although the original meta-data does not contain the query terms. If the document turns out to be indeed relevant for the user, the provided search result snippet may feature interesting extra information about the referred entity. If the document is not relevant, the user may identify the false paradigmatic relation in the snippet and is saved from following false assumptions. For this reason, we propose to progress from the expansion paradigm in this paper and propose a new paradigm for improving meta-data retrieval: pseudo descriptions.

In contrast to the expansion paradigm, the goal of pseudo descriptions is to compile for each meta-data document d a cohesive text field p which (1) is a collected but limited set of linguistic units that are relevant for the entity referred to in the meta-data, and which (2) at the same time, encodes the paradigmatic term relations that are needed to alleviate the vocabulary mismatch and the headroom problem. Due to the cohesive text, in the best case, pseudo descriptions may even increase search effectiveness compared to expansion approaches since proximity features are now available for relevance scoring. From a search engine developer's point of view, pseudo descriptions can be handled just like any other text field. Advanced expansion technology is not required. Each pseudo description p is represented canonically to d as a vector $\mathbf{p}$, and a multi-field query is used for the relevance computation:

$$\mathrm{RSV}_p(q, d) = w_1\, \mathbf{q}^\intercal \mathbf{d} + w_2\, \mathbf{q}^\intercal \mathbf{p} \tag{2}$$

Pseudo-descriptions are, for the reasons mentioned above, appealing for meta-data search, but the question is if they can be produced in such a way that the two desired properties are met. To this end, in Section 3, we propose to exploit the search engine result page (SERP) that is returned when issuing meta-data as a search query to a reference search engine as an ad-hoc approach. Similar to pseudo-relevance feedback methods for query expansion, we assume that the SERP for the top documents represents a valid description of the meta-data's entity. To tailor the SERP to the approximation of important paradigmatic relations, we disclose a relevance dependent SERP generation strategy where the relevance score of a reference document controls the amount of linguistic units this document contributes to the SERP.

In Section 4, we report on a series of experiments we carry out to evaluate whether pseudo-descriptions can reach the effectiveness of document expansion approaches. The experiments are based on different TREC collections where we withhold the content information to simulate a meta-data retrieval scenario. As reference collections, the TREC collections themselves as well as Wikipedia and the CommonCrawl are used. Though, not surprisingly, retrieval with full content information is superior to both paradigms pseudo-descriptions and document expansion, our relevance dependent approach for the former paradigm is consistently en par with the document expansion baseline, and improves on the baseline occasionally. In Section 5, we conclude with a discussion of open research questions and possibilities to further improve the generation of pseudo descriptions.

2 RELATED WORK

Since we introduce pseudo-descriptions as an alternative paradigm to document expansion for search in meta-data, no existing approaches for their generation exist. Therefore, in this section, we point out relations to other information retrieval tasks, and review existing work on document expansion.

2.1 Related IR Tasks

For the generation of pseudo descriptions for a given meta-data collection, we propose a process in Section 3 that comprises two principle steps.

In the first step, a pool of linguistic units from which a pseudo description can be compiled has to be retrieved for each meta-data record from a reference collection. If the lexical information in the meta-data records are short, e.g. only titles or product names, this step constitutes a classical information retrieval task including the generation of search result snippets [1, 24]. If the lexical information is more verbose, care has to be taken to not include too many irrelevant terms into the query. In this case, this first step becomes related to query-by-example retrieval, where keyword extraction is commonly used to choose appropriate query terms for an example document [8, 17]. A further connection can be drawn to candidate retrieval for plagiarism detection, where a multitude of queries is commonly generated to retrieve potential source documents for different sections of a suspicious document [9]. If temporal information is available in the meta-data records, aspects of temporal information retrieval become relevant for this first step [10].

In the second step, a subset of the candidate pool has to be selected and compiled into a text field. This task is similar to multi-document summarization [14, 15]. However, the objective is not a coherent, redundant-free text but a text that, if turned into a vector representation, approximates the result of applying a document expansion approach. In Section 3, we present an algorithm tailored to this objective.

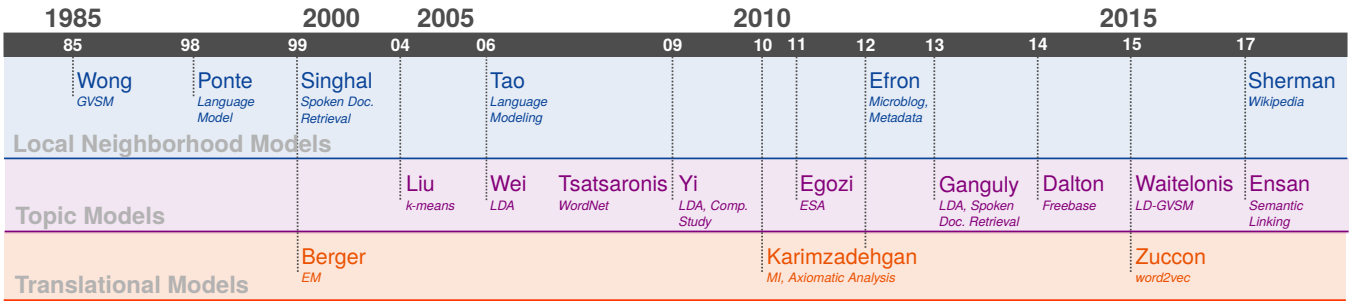

Figure 2: Document expansion approaches grouped by the problem perspective taken and arranged over time. Each publication is denoted by its first author and listed in the references. The keywords stated below the author outline characteristic properties of the respective work.

2.2 Document Expansion Review

To devise an algorithm for the generation of pseudo descriptions, it is helpful to know about the existing document expansion approaches. As outlined in Section 1, the goal of document expansion is to alleviate the vocabulary mismatch problem and the headroom problem by incorporating paradigmatic term relations as a vector $\mathbf{d}'$ into the relevance score computation. As one of the first attempts in this regard, the Generalized Vector Space Model (GVSM) [27] presented by Wong et al. in 1985 can be counted. The GVSM incorporates term relations by plugging a $\mathrm{n} \times \mathrm{n}$ term relation matrix $\mathbf{G}$ into the relevance function:

$$\mathrm{RSV}_{\mathrm{GVSM}}(q, d) = \mathbf{q}^\intercal \, \mathbf{G} \, \mathbf{d} \quad with \quad \mathbf{G} \in \mathbb{R}^{\mathrm{n} \times \mathrm{n}}$$

Each element $t_{ij} \in G$ denotes how strong the term t_j correlates with the term t_i. The GVSM can be interpreted as a document expansion approach, if $\mathbf{G}$ is first multiplied by $\mathbf{d}$ to obtain the expansion vector $\mathbf{d}' = \mathbf{G}\,\mathbf{d}$, which is then used in the relevance function $\mathrm{RSV}_{\mathbf{d}'}$ in Equation 1. Likewise, a query expansion approach arises when first multiplying $\mathbf{q}^\intercal$ with $\mathbf{G}$, which highlights the close relationship of the two expansion paradigms. Interestingly, document expansion papers commonly report increased retrieval performance when applying both query and document expansion. In terms of the GVSM, this corresponds to raising $\mathbf{G}$ to the power of two, which intuitively uncovers second-order term relations, i.e., terms that co-occur with the same terms (such as synonyms).

Since the publication of the GVSM, a large amount of document expansion approaches have been proposed until recently. An overview of the approaches is illustrated in Figure 2. Though most of the approaches are framed within the Language Model [18] by Ponte and Croft, we will show here that is possible to express most of them also within simple GVSM algebra. In the following paragraphs, we review document expansion approaches that exploit a lexical document collection. The key difference of the approaches is how the values in $\mathbf{G}$ are determined. Please note that other approaches exist that rely on structured knowledge bases such as Wordnet [23] or LinkedData [3, 6, 25]. However, it is not immediately clear how pseudo descriptions featuring a cohesive text can be obtained from these sources. Also note that, in order to keep the equations simple and general, normalization steps that are needed to exactly resemble the cited approaches are omitted in the following equations. In particular, every multiplication that is supposed to compute the cosine similarity of two vectors assumes that the vectors are L_2-normalized, first. For matrix rows, columns, and vectors that are supposed to be probability distributions, L_1-normalization is assumed.

In their original paper, Wong et al. propose to use term co-occurrence statistics obtained from the document collection $D = \{d_1, \ldots, d_m\}$ itself as estimates for the entries of $\mathbf{G}$. In follow-up work, also the use of external collections (later denoted as E) for this purpose, especially Wikipedia [5, 20], has been proposed. Co-occurrence statistics can be computed by multiplying the normalized term-document matrix $\mathbf{D}$ of D with its transpose [16], letting the computation of the document expansion vector become:

$$\mathbf{d}'_{\mathrm{cooc}} = \mathbf{D}\,\mathbf{D}^\intercal \, \mathbf{d} \quad with \quad \mathbf{D} \in \mathbb{R}^{\mathrm{n} \times \mathrm{m}}$$

As an alternative, computing the mutual information of the terms within D has been proposed in the context of translational document expansions models [11, 12].

A further avenue of research on document expansion applies topic modeling to D. Topic models represent documents as a distribution over k topics, which in turn are distributions over n terms. Early topic based document expansion approaches apply cluster analysis to $\mathbf{D}$ to obtain a set of k topic clusters, and then represent $\mathbf{d}$ by its cosine similarity to the cluster centroids [13]. If the k cluster centroids are stacked as a $\mathrm{n} \times \mathrm{k}$ matrix $\mathbf{C}$, document expansion with clustering can be expressed as:

$$\mathbf{d}'_{\mathrm{topic}} = \mathbf{C}\,\mathbf{C}^\intercal \, \mathbf{d} \quad with \quad \mathbf{C} \in \mathbb{R}^{\mathrm{n} \times \mathrm{k}}$$

Canonically, document expansion with the explicit topic model ESA [5] can be expressed. ESA treats articles from Wikipedia as explicit topics, and uses the vector representations of the article texts to compile a set of k topic vectors, which then make up the matrix $\mathbf{C}$ in the equation above. Furthermore, also document expansion with word embeddings [30] can be represented this way. In this case, $\mathbf{C}$ contains the embeddings for each of the n terms over the k latent dimensions. To compute related terms, $\mathbf{d}$ is first multiplied by $\mathbf{C}^\intercal$ to obtain a latent representation for $\mathbf{d}$, which is then normalized and multiplied by $\mathbf{C}$ to obtain a vector with cosine similarities of all n terms.

Though in principle, also document expansion with the latent topic model LDA [7, 26, 28] could be applied this way, the representation of documents under LDA is conventionally not determined by their similarity to the topic vectors but through statistical inference:

$$\mathbf{d}'_{\mathrm{LDA}} = \mathbf{C}\ \mathrm{lda}(\mathbf{C}, \mathbf{d}) \quad with \quad \mathbf{C} \in \mathbb{R}^{\mathrm{n} \times \mathrm{k}}$$

Figure 3: Mockup of a meta-data search engine with pseudo descriptions. Shown is the first result for TREC query 70, "surrogate motherhood". The result snippet features parts of the document's pseudo-description. The red boxes surround terms that appear in the document's title (=meta-data).

In the equation above, $\text{lda}(\mathbf{C}, \mathbf{d})$ is a function that infers a representation of $\mathbf{d}$ within the latent topic space spanned by $\mathbf{C}$. A further function useful to express the final stream of research on document expansion reviewed in this section is the function $\text{top}_\text{r}(\cdot)$, which sets all but the r highest values of a vector to zero.

The function $\text{top}_\text{r}(\cdot)$ is needed in the context of local neighborhood document expansion models [4, 20, 22]. Approaches of these kind interpret the multiplication of $\mathbf{d}$ with the term-document matrix $\mathbf{D}^\intercal$ as a retrieval step where $\mathbf{d}$ serves as search query which is issued against D. The result is a vector of relevance scores over D. While co-occurrence models use all relevance scores to compute the final expansion vector, local neighborhood models pick only the r most relevant documents:

$$\mathbf{d'}_{\text{local}} = \mathbf{D}\,\text{top}_\text{r}(\mathbf{D}^\intercal\,\mathbf{d}) \quad \textit{with} \quad \mathbf{D} \in \mathbb{R}^{\text{n}\times\text{m}}$$

With regards to the generation of pseudo descriptions discussed in the next section, the local neighborhood approach is conceptually closest to the approach we present for the generation of pseudo descriptions. We therefor use this approach as our baseline.

3 APPROACH

Pseudo-descriptions constitute an alternative paradigm to document expansion for alleviating the vocabulary mismatch and the headroom problem in the context of meta-data retrieval. In contrast to document expansion approaches, not a term vector $\mathbf{d'}$ is computed that encodes paradigmatic term relations, but term relations are encoded as an additional linguistic meta-data field p (see Equation 2 in Section 1.3). The main benefit of such a field is that the linguistic units that give rise to assume certain paradigmatic term relations are available to the meta-search engine for result snippet generation. This benefit is illustrated in Figure 3, which shows a search result for one of the TREC queries used in our experiments. Though the title of the retrieved meta-data document does not feature any of the query terms, its pseudo description features the query terms, allowing the search engine (1) to conclude that the document is relevant, and (2) to make the evidence used for this conclusion transparent for the user (in Figure 3, bold terms in the pseudo description excerpt indicate query term matches).

The question addressed in the remainder of this section is how pseudo descriptions should be computed for a meta-data collection D on the basis of an external reference collection E. As already mentioned in Section 2.1, we propose to model the generation of pseudo descriptions as a two step process: a candidate retrieval and a candidate selection process.

3.1 Candidate Retrieval

In the first step, linguistic units that are relevant for a meta-data record $d \in D$ have to be retrieved from an external reference collection E. Being in an information retrieval context, a natural way to obtain relevant linguistic units is to employ a search engine indexing the reference collection for this purpose. Relevant linguistic units are obtained by formulating a search query on the basis of the meta-data record, submitting this search query to the reference search engine, and taking the document snippets of the returned search result page (SERP) as the pool S of relevant linguistic units:

$$S = \text{serp}_{l,u}(\text{E}, \text{top}_\text{r}(\mathbf{E}^\intercal\,\mathbf{d})) \quad \textit{with} \quad \mathbf{E} \in \mathbb{R}^{\text{n}\times\text{m}}$$

In the equation above, the meta-data query of d, denoted as $\mathbf{d}$, is first multiplied by $\mathbf{E}^\intercal$ to obtain the relevance scores for all reference documents in E. Analogous to local neighborhood document expansion, the function $\text{top}_\text{r}(\cdot)$ is then applied to filter out all but the r most relevant documents. Instead of now aggregating the top r document vectors relative to their relevance scores to obtain an expansion vector $\mathbf{d'}$, we use the SERP-generation function $\text{serp}(\cdot)$ of the reference search engine to obtain relevant linguistic units on sub-document level. By default, the reference search engine we employ for our experiments, Elastic Search, uses the Lucene Unified highlighter, which "breaks the text into sentences and uses the BM25 algorithm to score individual sentences as if they were documents in the corpus".[1] The parameters l and u of the serp function specify the target length of a single result snippet and the target amount of snippets that should be returned for each relevant document, respectively. As a consequence of this snippet generation approach, the linguistic units obtained feature subsets of the meta-data query terms within the context of one or a few sentences, making these linguistic units very appropriate for our scenario. In Figure 3, the meta-data query terms are surrounded by red boxes to highlight how the pseudo description connects the user's search query ("surrogate motherhood") with the meta-data title ("Judge Accepts Baby M Settlement"). Note that, by just joining all of the (up to) $r * u$ returned linguistic units $S = \{s_{1,1}, \ldots, s_{r,u}\}$ together, a pseudo description $p_{\text{adhoc}} = \text{join}(S)$ can be produced already after this first step. The resulting vector representation of p_{adhoc} is the sum over the vector representations $s_{i,j}$ of all linguistic units in S:

$$\mathbf{p}_{\text{adhoc}} = \sum_{i=1}^{r}\sum_{j=1}^{u_i} \mathbf{s_{i,j}}$$

We refer to this one-step approach as *ad-hoc* approach in our experiments, where we request $u = 1$ snippet with a target size of $l = 250$ characters for the top $r = 10$ reference documents.

3.2 Candidate Selection

A limitation of the ad-hoc approach is that the generated pseudo descriptions p_{adhoc} are not tailored to approximate a particular

[1] https://www.elastic.co/guide/en/elasticsearch/reference/current/search-request-highlighting.html#unified-highlighter

document expansion vector $\mathbf{d'}$, and hence may misrepresent existing paradigmatic term relations. To remedy this short coming, we propose a candidate selection procedure as the second step of the pseudo description generation. To derive a first algorithm for this second step, we choose local neighborhood document expansion as our reference document expansion approach that we try to approximate. To recap, for a document d, local neighborhood document expansion first issues d as a search query to a search engine indexing the reference collection E. The vector representations $\mathbf{e}$ of the r most relevant documents in E are then summed up proportional to their relevance score $\text{RSV}(d, e)$. This can be, now including a normalization parameter z, written as:

$$\mathbf{d'}_{\text{local}} = \mathbf{E}\,\text{top}_{\text{r}}(\mathbf{E}^{\intercal}\,\mathbf{d}) = \sum_{e_i \in \text{top}_{\text{r}}} \frac{\text{RSV}(d, e_i)}{z}\,\mathbf{e_i} \tag{3}$$

To approximate the vector $\mathbf{d'}_{\text{local}}$ with a pseudo description, we make the simplifying assumptions that (1) the vector representation of every linguistic unit $s_{i,j} \in S$ is, respecting its relative size $l_{i,j}/|e_i|$, a fraction of the vector representation $\mathbf{e_i}$ from which the linguistic unit is taken, and that (2) every linguistic unit has the same length l. Under these assumptions, the vector representation of p_{adhoc} can be rewritten as:

$$\mathbf{p}_{\text{adhoc}} = \sum_{i=1}^{r}\sum_{j=1}^{u_i} \mathbf{s_{i,j}} \overset{(1)}{\approx} \sum_{i=1}^{r}\sum_{j=1}^{u_i} \frac{l_{i,j}}{|e_i|}\mathbf{e_i} \overset{(2)}{=} \sum_{i=1}^{r} u_i \frac{l}{|e_i|}\mathbf{e_i}$$

Comparing the vector representations $\mathbf{d'}_{\text{local}}$ with $\mathbf{p}_{\text{adhoc}}$, one can see that both vectors are sums over the r document vectors $\mathbf{e_i}$, and that the vectors would be equal if the fractions in the sums are equal, i.e., if $\frac{\text{RSV}(d,e_i)}{z} = u_i \frac{l}{|e_i|}$. Since the number of snippets u_i taken from each reference document e_i for inclusion into a pseudo description can be controlled, we propose, towards the generation of a pseudo description p^* that is tailored to local neighborhood document expansion, to modify u_i such that the fractions above become equal:

$$u_i^* = \left\lceil \frac{\text{RSV}(d, e_i)}{z}\,\frac{|e_i|}{l} \right\rceil$$

In the equation, the brackets denote rounding to the next integer, which is required since the number of snippets taken for a reference document must be a whole number. By using u_i^*, the final algorithm for our candidate selection algorithm, which we refer to as "relevance dependent", can be written as follows:

$$p^* = \text{join}(S^*) = \text{join}\left(\bigcup_{i=1}^{r}\bigcup_{j=1}^{u_i^*} s_{i,j}\right)$$

A final point to discuss is the normalization parameter z. Since local neighborhood document expansion has been proposed in the context of the Language Model, where any expansion vectors $\mathbf{d'}$ are supposed to be probability distributions (L_1 normalized), the parameter z is commonly chosen to be $z = \sum_{e_i \in \text{top}_{\text{r}}} \text{RSV}(d, e_i)$. However, we believe that this value is not appropriate in our context for the following reason: If for a document d only one marginally relevant document can be retrieved from the reference collection E, a L_1-normalizing value for z would have the effect that all of the snippets of this document end up in the pseudo description, likely having a negative effect on the overall retrieval performance. Conversely, if many highly relevant documents are retrieved, only few snippets would be taken from each of these documents, and significant paradigmatic term relations may be missed. Therefor, we propose to set z to be the maximum retrieval score that could be expected for the query representation of a meta-data document. This way, any marginally relevant document also contributes only marginally to a pseudo description, whereas any highly relevant document contributes many snippets. As the maximum expected value for a query highly dependents on the retrieval model of the reference search engine, which might in detail be unknown, in our experiments, we set z to the maximum retrieval score that we observed for all queries submitted to a reference search engine.

4 EVALUATION

In this section, we report on a series of experiments we conducted to compare the retrieval performance of relevance dependent pseudo descriptions p^* with the performance of both ad-hoc pseudo descriptions p_{adhoc} and local neighborhood document expansions $\mathbf{d'}_{\text{local}}$. Since the goal of p^* is to approximate $\mathbf{d'}_{\text{local}}$, achieving performance characteristics en par with $\mathbf{d'}_{\text{local}}$ constitutes a positive evaluation result. Due to the availability of term proximity features when using pseudo descriptions, in the best case, small performance improvements can be anticipated.

4.1 Datasets

We use two popular information retrieval benchmark collections for our experiments. The TREC AP collection (242,917 Associated Press news articles) and the combined collections on TREC DISKS 4&5 (472,521 news articles from Financial Times Limited, Foreign Broadcast Information Service, and Los Angeles Times; excluded due to missing title field: Congressional Records and Federal Register). To simulate a meta-data retrieval scenario, only the title field of the collections are made available to the meta-data search engines. To measure retrieval performance, we issue short and long versions of TREC queries 51-150 against the AP collection, and of TREC queries 351-450 against the TREC DISKS 4&5 collection.

As reference collections for the generation of pseudo descriptions and document expansion vectors, we reuse the two TREC collections in two variants: In variant one, "without self", the complete collection is available but the document under consideration is in each case ignored when retrieved. In variant two, "without titles", only the content field of the collections is available to simulate retrieval in an unstructured text collection with at least one highly relevant document. Furthermore, Wikipedia (2,104,323 articles)[2] is used as an encyclopedic reference collection, as well as the CommonCrawl (1,630,502,843 web pages)[3] to simulate, with more than a billion web pages, retrieval against the Web.

4.2 Experimental Setup

To compute the retrieval performance of the three approaches $\mathbf{d'}_{\text{local}}$, p_{adhoc}, and p^* across all combinations of meta-data and reference collections, we proceed in three steps.

[2]Not considered were lists, disambiguations, and short (#words < 250) articles.
[3]http://commoncrawl.org/2015/12/november-2015-crawl-archive-now-available/

Table 1: Evaluation results of the proposed approaches $\mathbf{d'}_{\text{local}}$, p_{adhoc}, and p^* for two meta-data collections and four reference collections. Reported are nDCG@20 and MAP performance scores for the available short and long query versions. $\mathbf{d'}_{\text{local}}$ is regarded as baseline, and ↑ / ↓ indicate improvement / decline over the baseline. † / ‡ indicate statistical significance of a result according to the Wilcoxon signed-rank test at the levels $p < 0.05$ / $p < 0.01$, respectively.

Collection	Reference Collection	Approach	nDCG@20 short	nDCG@20 long	MAP short	MAP long
AP	-	-	0.405	0.516	0.226	0.298
TREC AP Titles	-	-	0.255	0.277	0.055	0.067
	AP without self	$\mathbf{d'}_{\text{local}}$	0.329	0.389	**0.151**	**0.173**
		p_{adhoc}	$0.319^{\downarrow}$	$0.381^{\downarrow}$	$0.108^{\downarrow}$	$0.137^{\downarrow}$
		p^*	$\mathbf{0.335}^{\uparrow}$	$\mathbf{0.418}^{\uparrow}$	$0.128^{\downarrow}$	$0.166^{\downarrow}$
	AP without titles	$\mathbf{d'}_{\text{local}}$	0.327	0.391	0.148	0.171
		p_{adhoc}	$0.332^{\uparrow}$	$0.384^{\downarrow}$	$0.113^{\downarrow}$	$0.140^{\downarrow}$
		p^*	$\mathbf{0.356}^{\uparrow}$	$\mathbf{0.435}^{\uparrow}$	$\mathbf{0.150}^{\uparrow}$	$\mathbf{0.184}^{\uparrow}$
	Wikipedia	$\mathbf{d'}_{\text{local}}$	0.301	0.348	**0.103**	**0.121**
		p_{adhoc}	$0.299^{\downarrow}$	0.348	$0.082^{\downarrow}$	$0.103^{\downarrow}$
		p^*	$\mathbf{0.315}^{\uparrow}$	$\mathbf{0.361}^{\uparrow}$	$0.099^{\downarrow}$	$0.119^{\downarrow}$
	Common-Crawl	$\mathbf{d'}_{\text{local}}$	**0.330**	**0.397**	**0.112**	**0.135**
		p_{adhoc}	$0.299^{\downarrow}$	$0.355^{\downarrow}$	$0.086^{\downarrow}$	$0.107^{\downarrow}$
		p^*	**0.330**	$0.382^{\downarrow}$	$0.103^{\downarrow}$	$0.126^{\downarrow}$

Collection	Reference Collection	Approach	nDCG@20 short	nDCG@20 long	MAP short	MAP long
DISKS 4&5	-	-	0.426	0.492	0.217	0.261
TREC DISKS 4&5 Titles	-	-	0.217	0.222	0.060	0.060
	DISKS 4&5 without self	$\mathbf{d'}_{\text{local}}$	0.313	**0.354**	**0.130**	**0.145**
		p_{adhoc}	$0.291^{\downarrow}$	$0.329^{\downarrow}$	$0.094^{\downarrow\ddagger}$	$0.109^{\downarrow}$
		p^*	$\mathbf{0.314}^{\uparrow\dagger}$	$0.353^{\downarrow}$	$0.110^{\downarrow\ddagger}$	$0.130^{\downarrow}$
	DISKS 4&5 without titles	$\mathbf{d'}_{\text{local}}$	0.304	0.350	0.121	0.137
		p_{adhoc}	$0.291^{\downarrow\dagger}$	$0.329^{\downarrow}$	$0.093^{\downarrow\ddagger}$	$0.109^{\downarrow}$
		p^*	$\mathbf{0.333}^{\uparrow\dagger}$	$\mathbf{0.379}^{\uparrow}$	$\mathbf{0.126}^{\uparrow\ddagger}$	$\mathbf{0.149}^{\uparrow}$
	Wikipedia	$\mathbf{d'}_{\text{local}}$	0.286	0.319	**0.100**	0.111
		p_{adhoc}	$0.296^{\uparrow\dagger}$	$0.306^{\downarrow}$	$0.086^{\downarrow\ddagger}$	$0.095^{\downarrow}$
		p^*	$\mathbf{0.308}^{\uparrow\dagger}$	$\mathbf{0.332}^{\uparrow}$	$0.098^{\downarrow\ddagger}$	$\mathbf{0.113}^{\uparrow}$
	Common-Crawl	$\mathbf{d'}_{\text{local}}$	**0.313**	**0.372**	**0.111**	**0.131**
		p_{adhoc}	$0.280^{\downarrow\dagger}$	$0.323^{\downarrow}$	$0.080^{\downarrow}$	$0.097^{\downarrow}$
		p^*	$0.302^{\downarrow\dagger}$	$0.357^{\downarrow}$	$0.098^{\downarrow}$	$0.121^{\downarrow}$

Step 1. In the first step, all six reference collections (2 x TREC AP, 2x TREC DISKS 4&5, Wikipedia, and CommonCrawl) are indexed using the prevalent open source search engine Elastic Search with default settings (~BM25 retrieval model).

Step 2. In the second step, the titles of both meta-data collections are issued, without stopwords, as queries to each of the reference collections. To generate the expansion vectors $\mathbf{d'}_{\text{local}}$, the best performing settings reported by Tao et al. [22] are adopted, i.e., the term vectors of the $r = 100$ most relevant documents were fetched for each query and summed up relative to their retrieval score using the L_1-normalizing version for the parameter z (cf. Equation 3). To generate the ad-hoc pseudo descriptions p_{adhoc}, $u = 1$ document snippet is requested from Elastic Search with a target size of $l = 250$ characters for the top $r = 10$ reference documents. To generate the relevance dependent pseudo descriptions p^*, $u = 10$ candidate document snippets are initially requested from Elastic Search, also with a target size of $l = 250$ characters for the top $r = 10$ reference documents. If the relevance dependent candidate selection algorithm suggests to add more snippets than returned for a document, all available snippets are just added to p^* without any compensation.

Step 3. In the third step, the generated pseudo descriptions and document expansions are separately added as additional field to the respective meta-data collection to yield "enriched" meta-data collections. For each of the two meta-data collections, three approaches have been applied to six different reference collections, giving a total of 36 enriched meta-data collections. Note that to keep the evaluation table concise, we do not report on the results across meta-data collections in this paper, e.g., not on results obtained for TREC AP meta-data with TREC DISKS 4&5 as the reference collection. Each enriched meta-data collections is again indexed with Elastic Search. To be able to process the document expansion vectors $\mathbf{d'}_{\text{local}}$ properly, we implemented a custom plugin for Elastic Search.[4] For the pseudo descriptions, no custom code is required. To evaluate the retrieval performance of the enriched meta-data collections, the (short and long) TREC queries available for each meta-data collection are issued as multi-match query against the respective enriched meta-data search engines, and the resulting ranking is evaluated with the `trec_eval` script. Multi-match queries allow to combine retrieval scores obtained from multiple fields. In our case, retrieval scores for the title field as well as the added "enriched" field are combined using different field weights w_1 and w_2 (cf. Equations 1 and 2 in Section 1). To find optimal weights for every enriched meta-data collection, we set $w_1 = 1 - \lambda$ and $w_2 = \lambda$ (Jelinek-Mercer smoothing [29]), and performed a grid search over λ.

4.3 Overall Results

The performance results of all approaches under optimal λ values are summarized in Table 1. On the left, results for the TREC AP collection are shown, the results for TREC DISKS 4&5 are shown on the right. As primary retrieval performance measure, nDCG@20 is reported, which has been found to correlate better than other measures with user preferences [19]. In addition, performance in terms of MAP is reported. In the first row with results, the retrieval performance of a search engine indexing the full TREC collections are given as reference. In both cases, this (practically not available) search engine achieves the best performance, indicating that there is still room for improvement. Conversely, the second result row shows how a search engine indexing only the title information performs. In both cases, the title only search engine performs worst, indicating that every of the evaluated approaches is worth applying.

Comparing the performance of the three evaluated approaches among each other, it becomes apparent that the ad-hoc pseudo

[4] Available at https://github.com/nadre/elasticsearch-delimited-tf-token-filter.

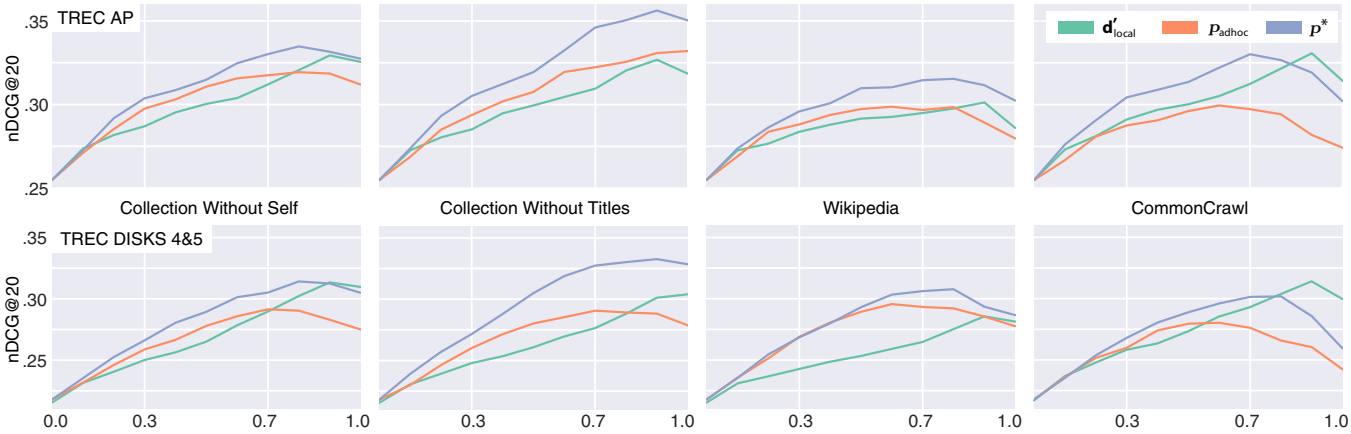

Figure 4: Short query performance of the evaluated approaches for different values of $\lambda \in [0, 1]$. Each plot shows a different combination of meta-data and reference collection.

descriptions p_{adhoc} commonly perform slightly worse than the other two approaches. More importantly, the main finding of the experiment is that the performance of the relevance dependent pseudo descriptions p^* is indeed en par with the performance of the local neighborhood document expansions $\mathbf{d}'_{\text{local}}$. Except for the CommonCrawl reference collection (and one further result), the nDCG@20 performance is even slightly better.

4.4 nDCG vs. MAP

Studying the MAP performance scores in Table 1, one can observe that in 11 of the 16 cases, our approach p^* performs slightly worse than the document expansion approach $\mathbf{d}'_{\text{local}}$ for this performance measure. Compared to nDCG@20, MAP puts more emphasis on retrieval recall then on retrieval precision, giving rise to the hypothesis that our approach misses to encode some of the weak paradigmatic term relations needed to detect relevant documents with high recall. A possible reason for this is that fewer linguistic units are encoded in the vector representation of p^*. However, in most practical settings, a high retrieval precision as emphasized by nDCG@20 increases the user experience more than a high retrieval recall, and hence we argue that it is worth trading a small gain in nDCG@20 against a small loss in terms of MAP.

4.5 Influence of Reference Collections

Taking a closer look at the nDCG@20 performance changes under different reference collections, two observations are striking. First, our approach p^* performs better under the internal collections "without titles" than on the internal collections "without self", whereas the opposite is true for the document expansion approach $\mathbf{d}'_{\text{local}}$. We hypothesize that this phenomenon arises due to the different normalization parameters used in the two approaches. The maximum normalization used in our approach ensures that many snippets of the original document contained in the "without title" collections are included in p^*. For $\mathbf{d}'_{\text{local}}$, the influence of the original document is, because of the L_1 normalization, comparably low, preventing its full exploitation. The second interesting observation is that our approach is able to exploit Wikipedia better than $\mathbf{d}'_{\text{local}}$ does, whereas the opposite is true for the CommonCrawl. Our hypothesis here is that, in the case of Wikipedia, only small subsections of the retrieved documents really talk about the topic of the meta-data document, while the article itself is about a more general topic. Since, in contrast to the document expansions, pseudo descriptions consider linguistic units on sub-document level, it is possible that just these relevant subsections are taken. In the case of the CommonCrawl, we hypothesize that many relevant articles appear in this collection. Since $\mathbf{d}'_{\text{local}}$ takes more documents into account, it is likely that here, paradigmatic term relations can be inferred more robustly by this approach. To further investigate this phenomenon, a systematic exploration of the approaches' parameters is required, which is, however, beyond the paper's scope.

4.6 Lambda Plots

As noted in the experimental setup, the optimal value for the smoothing parameter λ is determined via grid search. In Figure 4, the performance development over λ is illustrated for the three evaluated approaches. Each plot refers to a specific combination of meta-data and reference collection. From the plots, it can be observed that the optimal λ is typically between 0.7 and 1.0 for the two best approaches. Further, across the value range, p^* is consistently better performing than the other approaches. Only with CommonCrawl as the reference collection, $\mathbf{d}'_{\text{local}}$ achieves better performances near the end of the value range. Towards the validity of our approach, this observation substantiates our belief that relevance dependent pseudo descriptions are a robust alternative for local neighborhood document expansions.

4.7 Benefit of Term Proximities

One of the benefits of pseudo descriptions is that term proximity features can be used by the meta-data search engine for the relevance computation. As a final evaluation, in Table 2, the differences in retrieval performance when toggling the term proximity feature of Elastic Search on and off are presented for the full content search engine and the search engines indexing the pseudo descriptions p^*. In all but one case, using the proximity feature leads to slight retrieval performance increases. For the TREC DISKS 4&5 collection, the improvements are significant at the $p < 0.05$ level

Table 2: nDCG@20 effectiveness comparison between disabled and enabled term proximity ranking. ↑ / ↓ denotes improvement / decline and † indicates statistical significance at the $p < 0.05$ level, according to Wilcoxon signed-rank test.

Collection	Proximity	Title and Text	p^*			
			Coll. w/o self	Coll. w/o titles	Wiki	Common-crawl
AP	Off	0.402	0.335	0.349	0.316	0.324
	On	$0.405^{\uparrow}$	$0.335^{\uparrow}$	$0.356^{\uparrow}$	$0.315^{\downarrow}$	$0.330^{\uparrow}$
DISKS 4&5	Off	0.420	0.309	0.324	0.305	0.298
	On	$0.426^{\uparrow\dagger}$	$0.314^{\uparrow\dagger}$	$0.333^{\uparrow\dagger}$	$0.308^{\uparrow\dagger}$	$0.302^{\uparrow\dagger}$

according to Wilcoxon signed-rank test. Again, we see this result as a substantiation of the validity of the pseudo description paradigm.

5 CONCLUSION

Main contribution of this paper is the introduction of pseudo descriptions as a new paradigm for improving search in meta-data. Towards conceptual and operational models for the generation of pseudo descriptions, we first reviewed existing document expansion approaches in terms of the Generalized Vector Space Model. Based on the gained insights, a general two-step procedure for the generation of meta-data description is presented, and we propose an algorithm that generates pseudo descriptions on the basis of search result pages while approximating local neighborhood document expansion. Related research questions include how non-lexical meta-data fields could be exploited, whether linguistic units other than search result snippets can be used as basis for pseudo descriptions, how other document expansion approaches such as topic or translational models can be approximated, and whether local neighborhood document expansion can be approximated with less strong assumptions. The validity of our relevance dependent pseudo description approach is evaluated in Section 4, where we find that our approach shows performance characteristics en par with local neighborhood document expansion. An open research question in this regard is how beneficial the presentation of pseudo descriptions in search results is in practice. Though we observed a multitude of promising result snippets produced from pseudo descriptions in our experiments (like the one in Figure 3), a scientific answer to this question requires a sophisticated user study.

REFERENCES

[1] Hannah Bast and Marjan Celikik. 2014. Efficient Index-Based Snippet Generation. *ACM Trans. Inf. Syst.* 32, 2, Article 6 (April 2014), 24 pages.

[2] Adam Berger and John Lafferty. 1999. Information Retrieval As Statistical Translation. In *Proc. of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, NY, USA, 222–229.

[3] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity Query Feature Expansion Using Knowledge Base Links. In *Proc. of the 37th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '14)*. ACM, 365–374.

[4] Miles Efron, Peter Organisciak, and Katrina Fenlon. 2012. Improving Retrieval of Short Texts Through Document Expansion. In *Proc. of the 35th Int. ACM SIGIR Conf. on Research and Dev. in Information Retrieval (SIGIR '12)*. ACM, 911–920.

[5] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. 2011. Concept-Based Information Retrieval Using Explicit Semantic Analysis. *ACM Trans. Inf. Syst.* 29, 2, Article 8 (April 2011), 34 pages.

[6] Faezeh Ensan and Ebrahim Bagheri. 2017. Document Retrieval Model Through Semantic Linking. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, NY, USA, 181–190.

[7] Debasis Ganguly, Johannes Leveling, and Gareth J.F. Jones. 2013. An LDA-smoothed Relevance Model for Document Expansion: A Case Study for Spoken Document Retrieval. In *Proc. of the 36th Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, NY, USA, 1057–1060.

[8] Matthias Hagen, Anna Beyer, Tim Gollub, Kristof Komlossy, and Benno Stein. 2016. Supporting Scholarly Search with Keyqueries. In *Advances in Information Retrieval. 38th European Conf. on IR Research (ECIR 16)* Springer, 507–520.

[9] Matthias Hagen, Martin Potthast, Payam Adineh, Ehsan Fatehifar, and Benno Stein. 2017. Source Retrieval for Web-Scale Text Reuse Detection. In *Proc. of the 26th ACM Int. Conf. on Inf. and Know. Management (CIKM 17)*. ACM, 2091–2094.

[10] Nattiya Kanhabua, Roi Blanco, and Kjetil NÃ¿rvÃ¦g. 2015. Temporal Information Retrieval. *Foundations and TrendsÂő in Information Retrieval* 9, 2 (2015), 91–208.

[11] Maryam Karimzadehgan and ChengXiang Zhai. 2010. Estimation of Statistical Translation Models Based on Mutual Information for Ad Hoc Information Retrieval. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM, NY, USA, 323–330.

[12] Maryam Karimzadehgan and ChengXiang Zhai. 2012. Axiomatic Analysis of Translation Language Model for Information Retrieval. In *Proc. of the 34th Europ. Conf. on Advances in Information Retrieval (ECIR'12)*. Springer-Verlag, 268–280.

[13] Xiaoyong Liu and W. Bruce Croft. 2004. Cluster-based Retrieval Using Language Models. In *Proceedings of the 27th Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '04)*. ACM, NY, USA, 186–193.

[14] Shulei Ma, Zhi-Hong Deng, and Yunlun Yang. 2016. An Unsupervised Multi-Document Summarization Framework Based on Neural Document Model. In *COLING*.

[15] Kaustubh Mani, Ishan Verma, and Lipika Dey. 2017. Multi-Document Summarization using Distributed Bag-of-Words Model. *CoRR* abs/1710.02745 (2017).

[16] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, NY, USA.

[17] Cristiano Nascimento, Alberto H.F. Laender, Altigran S. da Silva, and Marcos André Gonçalves. 2011. A Source Independent Framework for Research Paper Recommendation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*. ACM, NY, USA, 297–306.

[18] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *Proc. of the 21st Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '98)*. ACM, 275–281.

[19] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do User Preferences and Evaluation Measures Line Up?. In *Proc. of the 33rd Int. ACM SIGIR Conf. on Res. and Dev. in Inf. Ret. (SIGIR '10)*. ACM, 555–562.

[20] Garrick Sherman and Miles Efron. 2017. Document Expansion Using External Collections. In *Proc. of the 40th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '17)*. ACM, NY, USA, 1045–1048.

[21] Amit Singhal and Fernando Pereira. 1999. Document Expansion for Speech Retrieval. In *Proc. of the 22Nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '99)*. ACM, NY, USA, 34–41.

[22] Tao Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. 2006. Language Model Information Retrieval with Document Expansion. In *Proc of the Human Language Technology Conf. of the North American Chapter of the Ass. of Comp. Ling. (HLT-NAACL '06)*. Association for Computational Linguistics, 407–414.

[23] George Tsatsaronis and Vicky Panagiotopoulou. 2009. A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness. In *Proc. of the 12th Conf. of the Europ. Chapter of the Ass. for Computational Linguistics: Student Research Workshop (EACL '09)*. Association for Computational Linguistics, 70–78.

[24] Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E. Williams. 2007. Fast Generation of Result Snippets in Web Search. In *Proc. of the 30th Int. ACM SIGIR Conf. on Res. and Dev. in Information Retrieval (SIGIR '07)*. ACM, 127–134.

[25] Jörg Waitelonis, Claudia Exeler, and Harald Sack. 2015. Enabled Generalized Vector Space Model to Improve Document Retrieval. In *NLP-DBPEDIA@ISWC*.

[26] Xing Wei and W. Bruce Croft. 2006. LDA-based Document Models for Ad-hoc Retrieval. In *Proc. of the 29th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '06)*. ACM, NY, USA, 178–185.

[27] S. K. M. Wong, Wojciech Ziarko, and Patrick C. N. Wong. 1985. Generalized Vector Spaces Model in Information Retrieval. In *Proc. of the 8th Inter. ACM SIGIR Conf. on Res. and Dev. in Information Retrieval (SIGIR '85)*. ACM, 18–25.

[28] Xing Yi and James Allan. 2009. A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Proc. of the 31th European Conf. on IR Research on Advances in Information Retrieval (ECIR '09)*. Springer-Verlag, 29–41.

[29] Chengxiang Zhai and John Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proc. of the 24th Inter. ACM SIGIR Conf. on Res. and Dev. in IR (SIGIR '01)*. ACM, 334–342.

[30] Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. 2015. Integrating and Evaluating Neural Word Embeddings in Information Retrieval. In *Proc. of the 20th Aus. Doc. Comp. Symposium (ADCS '15)*. ACM, 8 pages.