Indiscriminateness in representation spaces of terms and documents

Vincent Claveau

▶ **To cite this version:**

Vincent Claveau. Indiscriminateness in representation spaces of terms and documents. ECIR 2018 - 40th European Conference in Information Retrieval, Mar 2018, Grenoble, France. pp.251-262, 10.1007/978-3-319-76941-7_19 . hal-01859568

HAL Id: hal-01859568
https://hal.science/hal-01859568

Submitted on 22 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Indiscriminateness in representation spaces of terms and documents

Vincent Claveau

Univ. Rennes, CNRS, IRISA
Campus de Beaulieu, Rennes, France
vincent.claveau@irisa.fr

Abstract. Examining the properties of representation spaces for documents or words in Information Retrieval (IR) – typically $\mathbb{R}^n$ with n large – brings precious insights to help the retrieval process. Recently, several authors have studied the real dimensionality of the datasets, called intrinsic dimensionality, in specific parts of these spaces [14]. They have shown that this dimensionality is chiefly tied with the notion of indiscriminateness among neighbors of a query point in the vector space. In this paper, we propose to revisit this notion in the specific case of IR. More precisely, we show how to estimate indiscriminateness from IR similarities in order to use it in representation spaces used for documents and words [18, 7]. We show that indiscriminateness may be used to characterize difficult queries; moreover we show that this notion, applied to word embeddings, can help to choose terms to use for query expansion.

Keywords: intrinsic dimensionality, indiscriminability, RSV scores, distributional thesauri, query expansion

1 Introduction

Examining the properties of representation spaces for documents or words in Information Retrieval (IR) – typically $\mathbb{R}^n$ with n large– brings precious insights to help the retrieval process. It is well-known that the dimensionality of the representation space is not the same as the dimensionality of the data. In the usual vector space model used in IR, the dimensionality of the representation space is the number of different words in the document collection, yet it often possible to represent the same documents in a space with much less dimensions. This fact is at the heart of techniques like *Latent Semantic Indexing* or *Latent Dirichlet Allocation* which reduce the dimensionality of the original (very sparse) vector space to a much smaller (and dense) space.

In this paper, we focus on the intrinsic dimensionality of the data, not from a global perspective (as for LSI or LDA), but more locally on portions of the space. For that purpose, we rely on the work of [14] and [2] which permit to define and estimate the local intrinsic dimensionality of the data (see Sec. 2). They showed that it can be used to measure the indiscriminateness of neighbors of a query, and thus to indirectly assess the potential quality of the answers to this query.

Since indiscriminateness depends on how the neighborhood is defined, and thus on the distance metric used, it is necessary to adapt it to RSV (*Relevance Status Value*) if one wants to use it in IR (Sect. 3). Then, we show in Sect. 4 how it can be used to analyze the representation space of documents in IR, for instance to detect difficult queries.

2 Related work

Characterizing the intrinsic dimensionality of data sets have been studied in different ways. For instance, embedding techniques or projection techniques build spaces with lower dimensionality in which data points are projected under certain conditions of discriminateness, like *Principal Component Analysis* (PCA), *Latent Semantic Indexing* (LSI), *Latent Dirichlet Allocation* (LDA) [8, 11] or *manifold learning* [22, 21, 26]. The intrinsic dimensionality of the whole data set is then the one of this new space obtained through projection.

Recently, [14] proposed a generalized expansion measure defining the local intrinsic dimension by examining how many points are met around a query point within a certain distance, and how it evolves when the distance augments. More formally, consider two balls centered in x_1 and x_2 with radius ϵ_1 et ϵ_2 in $\mathbb{R}^m$. The ratio between the volumes of these balls can be expressed as:

$$\frac{volume(B(x,\epsilon_1))}{volume(B(x,\epsilon_2))} = \left(\frac{\epsilon_1}{\epsilon_2}\right)^m$$

From that, one can define the dimensionality:

$$m = \frac{\ln(volume(B(x,\epsilon_1))) - \ln(volume(B(x,\epsilon_2)))}{\ln\epsilon_1 - \ln\epsilon_2}$$

The idea at the heart of the intrinsic dimension measure is to divert the previous equation by replacing the number of points in the volume instead of the volume itself [14, Sect IV B for justification]. Let us note $|B(x,\epsilon)|$ the number of points in $B(x,\epsilon)$; we have:

$$\hat{m} = \frac{\ln|(B(x,\epsilon_1)| - \ln|B(x,\epsilon_2)|}{\ln\epsilon_1 - \ln\epsilon_2}$$

The dimensionality is now the one of the data, not the one of the representation space. It is worth noting that this estimate is local to a point x (when considering $x = x_1 = x_2 = x$).

This intrinsic dimensionality model has been used in several ways for analyzing and building indexing structures for similarity search [4, 12, 13], and for anomaly detection [27]. Let us also cite the work of [3]; it does not rely on intrinsic dimensionality but the authors also exploit statistics on distance distribution in a similar context than ours in Sect. 4.

3 Use for Information Retrieval

The interest of intrinsic dimensionality for IR is its capacity to characterize the neighborhood of a query based on the documents surrounding it in the representation space. If this intrinsic dimensionality is very high, it means that a slight distance variation may completely change the set of documents that are considered as the closest to the query. Therefore, a high intrinsic dimensionality implies a high indiscriminateness of the documents around the query [14]. This is this exact property that we want to exploit here, provided that we can adapt it to the particular case of the similarity (Relevance Status Value, RSV) functions used in IR.

3.1 Limits

The intrinsic dimensionality definition previously given is set for a space in which the metric used is a distance, typically a L2 distance (Euclidean distance). It is used to define the set of points contained in the balls with different diameters (eg. for a ball centered in x with radius $r > 0$, the points d_i considered are those with $\mathrm{L2}(x, d_i) \leq r$). Yet, in IR, L2 is rarely used as RSV in the vector space model; instead, cosine has been widely used. As it has been shown [14], the intrinsic dimension definition can be used with such angular distances. The approach is the same as before: for a given vector, we compare the number of vectors in its neighborhood, that is, with angles lesser than ϵ_1 and ϵ_2.

Most of the common and modern RSV function can be written:

$$RSV(q, d) = \sum_{t \in q} w_q(t) \cdot w_d(t)$$

with $w_q(t)$ the weight of term t in query q and $w_d(t)$ the weight in document d, as illustrated in Tab. 1 (from [16]).

with the following notations:

$c(t, d)$	number of occurrences of term t in document d
$c(t, q)$	number of occurrences du term t in query q
N	number of documents in the collection
$df(t)$	number of documents containing term t
$dl(d)$	length of document d
$avdl$	average length of documents
$c(t, C)$	number of occurrences of term t in collection C
$p(t\|C)$	probability of term t for a language model of the collection

The RSV functions can be seen as simple scalar product between document vector d and query vector q, which we note $\langle q, d \rangle$. It differs from cosine in that it does not impose a L2 normalization of the vectors.

The absence of normalization is an important issue since it makes impossible to use the same principle as before to compute the intrinsic dimension. Indeed, for a query q, close documents (in terms of scalar product) may be at any L2 distance. More formally, for two thresholds values for the scalar product ϵ_1 and ϵ_2 ($\epsilon_1 \geq \epsilon_2$), the part of space containing points d_i such that $\epsilon_1 \geq \langle d_i, q \rangle \geq \epsilon_2$ is infinite, as illustrated in Fig. 3.

model	weighting
BM25+ $w_d(t)$	$\left(\frac{(k1+1)c(t,d)}{k1(1-b+b\cdot dl(d)/avdl)+c(t,d)} + \delta\right) \cdot \log \frac{N+1}{df(t)}$
BM25+ $w_q(t)$	$\frac{(k3+1)c(t,q)}{k3+c(t,q)}$
	with $k1, k_3, b$ and δ fixed parameters
PL2 $w_d(t)$	$\frac{tfn(t,d)\cdot\log_2(tfn(t,d)\cdot\lambda_t)+\log_2 e\cdot(1/\lambda_t - tfn(t,d))+0.5\log_2(2\pi\cdot tfn(t,d))}{tfn(t,d)+1}$
PL2 $w_q(t)$	$c(t,q)$
	with $tfn(t,d) = c(t,d)\cdot\log_2\left(1 + c\cdot\frac{avdl}{dl(d)}\right)$
	$c > 0$ a search parameter and $\lambda_t = \frac{N}{c(t,C)}$
Dir $w_d(t)$	$\log\left(\frac{\mu}{dl(d)+\mu} + \frac{c(t,d)}{(dl(d)+\mu)p(t\|C)}\right)$
Dir $w_q(t)$	$c(t,q)$
	$\mu > 0$ a smoothing parameter
Piv $w_d(t)$	$\frac{1+log(1+log(c(t,d)))}{1-s+s\cdot dl(d)/avdl} \cdot \log\frac{N+1}{df(t)}$ si $c(t,d) > 0$ and 0 else
Piv $w_q(t)$	$c(t,q)$
	with s a fixed parameter

Table 1: Weighting functions of terms in the query and the document for different state-of-the-art IR models: BM25+ [20, 16], Divergence From Randomness PL2 [1, 9], Language modeling with Dirichlet smoothing Dir [28], Pivoted Normalization Piv [23]

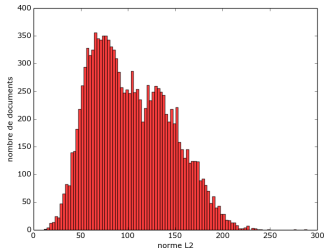

Fig. 1: Distribution of the L2 norms of documents in Tipster collection under BM25+ (modified version of BM25 proposed by [16])

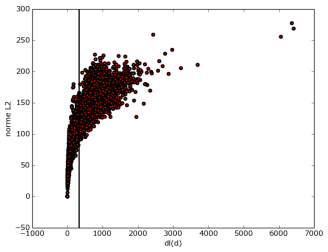

Fig. 2: L2 norms of documents in Tipster collection according to their length $dl(d)$ with BM25+; horizontal line is the average length ($avdl$)

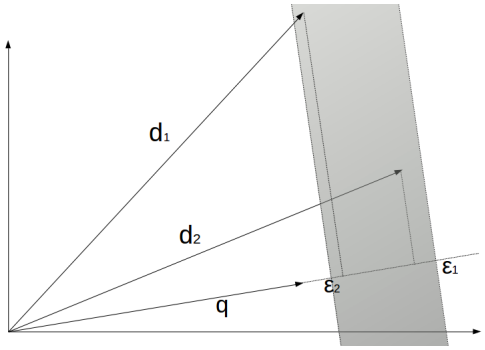

Fig. 3: In gray: portion of space defined by the set of points whose scalar products with a normed vector q lie between ϵ_1 and ϵ_2

3.2 Estimate with the Power Law exponent

In spite of the limit caused by the scalar product form of most RSV, we still aim at characterizing the intrinsic dimensionality, or at least the indiscriminateness, locally in the space. In previous work, [15, 2] has shown that the intrinsic dimensionality could be estimated from the repartition of the L2 distances between a query and the other points. In line with this, we propose to characterize indiscriminateness by examining the evolution of the number of neighbors (in our case, documents considered as close to the query) according to the RSV. This evolution can be interpreted as the repartition function of a random variable X which represents the RSV score between a given query and a document. More precisely, since we are only interested in the local behavior of X, that if to the closest documents, we only examine the repartition function for the highest RSV scores. The hypothesis we make is that the distribution of RSV can be locally modeled as a Power Law, that is:

$$f(x) = \lambda x^{-\alpha} \quad \text{with } \lambda \text{ a constant and} \quad \alpha > 1 \tag{1}$$

This is the exponent α which is characteristic of the indiscriminateness of the data. Formally, α cannot be linked to the intrinsic dimensionality as defined by [14] due to the problem raised by the use of scalar product as previously explained. In the IR case, we have n observations x_i, that is n RSV values of a query with its n closest neighbors (n highest RSV scores). It is thus possible to estimate α from these x_i. Among the various methods in the literature, the estimation based on log-likelihood has been shown to be the less biased [5]. Let the x_i be all the observations greater than a threshold x_{min}; α is then estimated as:

$$\hat{\alpha} = 1 + n \cdot \left(\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right)^{-1} \tag{2}$$

In the experiments reported below, we consider $n = 100$ observations x_i to estimate α, that is the 100 highest RSV scores (x_{min} is the 101st highest RSV score).

Fig. 4 shows the RSV score repartition as an histogram for a given query, and the corresponding Power Law whose exponent α is estimated as previously explained. Two facts are worth noting: first, the hypothesis we make about the distribution following a Power Law seems reasonable, since the histogram is typical for this distribution (the same observation holds for every query tested), and second, the estimation method is adequate, showing very few differences between the real distribution and the Power Law obtained with the estimated α.

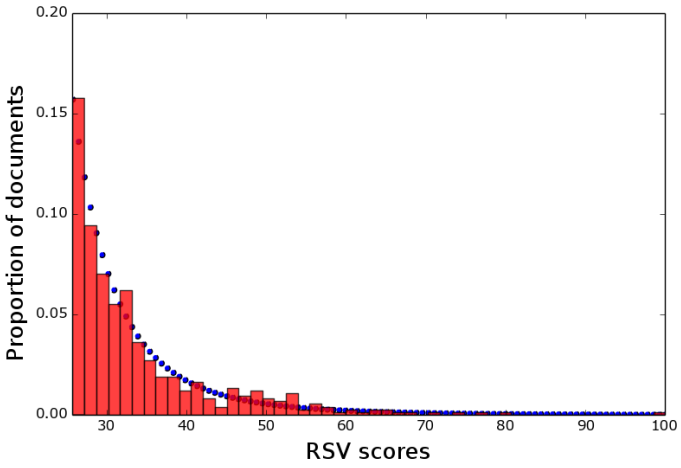

Fig. 4: Example RSV values repartition (red histogram) and the corresponding Power Law (blue) obtained with log-likelihood estimate of α from the RSV values

4 Experiments in the document space

In this section, we study how the indiscriminateness index α, as defined above, can be used in a standard IR framework. We show how α can be used to characterize the documents close to a query, either within the vector space model or in other spaces where the RSV can nonetheless be seen scalar products in Euclidean spaces.

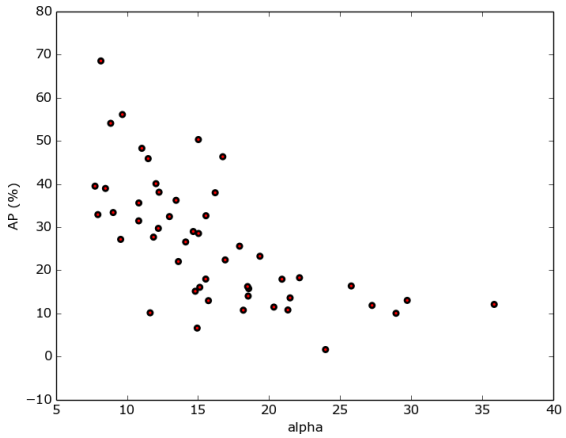

Fig. 5: Performance (AP) of queries from Tipster according to their index α with a BM25+ model

4.1 Data and evaluation scores

Two IR collections are used in our experiments: Tipster and OHSUMED [10]. Tipster contains more than 170 000 documents and 50 queries; it was used in TREC-2. The queries are composed of several parts, including the query itself and a narrative detailing the relevance criteria); in the experiments reported below, only the actual query part is used. OHSUMED contains 350,000 bibliographical notices from Medline and 106 queries from the TREC-9 filtering task. Performance are assessed with standard scores: Precision at different threshold (P@x), R-precision (R-prec), *Mean Average Precision* (MAP).

4.2 Detecting difficult queries

The distribution of documents around a query, or more precisely, the distribution of distances (or RSV) between a query and its closest documents can help characterize the difficulty of the query. In order to assess that, we look for a correlation between the index α (cf. Eqn. 2) around a query and the Average Precision (AP) of this query. This is illustrated in Fig. 5 on the Tipster collection, with BM25+ [16] as RSV.

The set of points exhibit the expected dependency between index α and the performance. In Tab. 2, we indicate the Pearson, Spearman and Kendall correlations (and their p-values) between the list of queries ordered by AP and the list of queries ordered by α on Tipster with a BM25+ model. The same information is given in Tab. 3 for OHSUMED with a Dirichlet LM (μ is set to 1000 for which it maximizes the MAP). The inverse correlation clearly appears: a low retrieval performance for query is related to a high indiscriminateness around this query.

coefficient	Value	p-value
Pearson r	-0.7150	$5.43e^{-09}$
Spearman ρ	-0.7753	$3.82e^{-11}$
Kendall τ	-0.5755	$3.69e^{-09}$

Table 2: Correlations (and their associated p-values) between AP and index α on Tipster with a BM25+ model

coefficient	Value	p-value
Pearson r	-0.4919	$9.85e^{-08}$
Spearman ρ	-0.6141	$3.26e^{-12}$
Kendall τ	-0.4494	$1.14e^{-11}$

Table 3: Correlations (and their associated p-values) between AP and index α on OHSUMED with a Dirichlet LM ($\mu = 1000$)

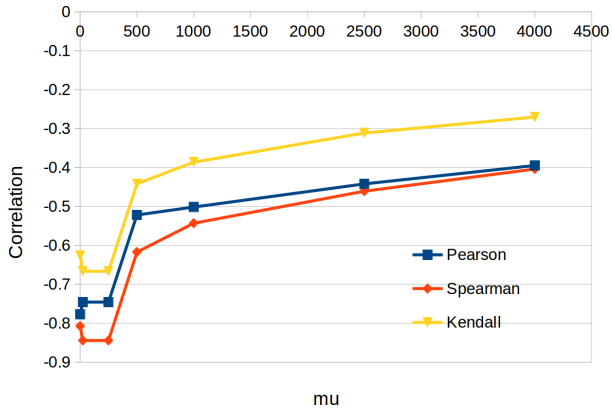

Fig. 6: Evolution of correlation scores (Pearson's r, Spearman's ρ, Kendall's τ) according to Dirichlet smoothing paramater μ.

Same observations also hold with other RSV functions on both collections, but some model parameters have a high impact on the results. For instance, we can observe the influence of smoothing on indiscriminateness within languages models. When smoothing is heavy (high μ in Dirichlet smoothing for example), documents tends to have similar weights for every query term, and thus similar RSV score, which thus makes the documents more difficult to discriminate (high α index). When μ is low, (inverse) correlation is high and tends to diminish when μ gets higher. It can be verified in Fig. 6, where the correlation between AP and α are given for several μ.

5 Query expansion

Projecting words in continuous representation spaces, such as vector spaces, has been widely studied recently. In these spaces, it is possible to find semantic proximity between words with the help vector-based distances. It makes it possible to build semantic lexicons in which each word is associated with its closest neighbors. In order to do so, these embedding techniques rely on the distributional hypothesis: close words share close contexts. By comparing contexts of words

(words occurring before and after), these techniques infer proximities between the words and/or a vector representation of the words. In this field, WORD2VEC [18] is very well-known: words are represented as vectors with the help of a neural approach. In recent work, [7] have shown that IR techniques could also be used to build such vector representation for words. In this later work, the authors use similarities like Okapi-BM25 between contexts of two words to build distributional thesauri. Finally, a word is represented by its similarity scores to every other word. This approach, called spectral, has yielded good results in numerous tasks [6].

Semantic lexicons can be used for query expansion: the closest semantic neighbors of query words are added to the query. It has been used as a way to evaluate the quality of the lexicons. In this section, we examine indiscriminateness, as previously defined, in representation spaces generated by WORD2VEC or the spectral approach. Since they also rely on similarity based on RSV, we use the same estimation technique of index α to characterize the local properties of the word spaces. In particular, we examine how α can be used in the query expansion task.

5.1 Framework

We adopt the following experimental framework. The IR collection is Tipster (see previous section). The distributional lexicons are the spectral one developed by [7] and a Word2Vec model trained on the GoogleNews (freely available at `https://code.google.com/p/word2vec/`). The IR system used is Indri [17, 24]. It is known to offer state-of-the-art performance, and moreover this probabilistic systems implements a combination of language modeling [19] and inference networks [25] which makes it possible to use operators AND OR... In the experiments report below, standard settings are used (Dirichlet smoothing parameter $\mu = 2500$). Thanks to its query language, this system allows us to easily expand the query with the semantic neighbors found in the distributional lexicon: the operator '`#syn`' aggregates counts of words considered as synonyms. In order to limit the effect of inflection (plural/singular) on the results, both plural and singular forms of the words are added to the query (original words of the query or words added from the lexicon). Performance is evaluated with the standard IR scores, by comparing results with and without expansion.

5.2 Experiments

Many uses for the α index of the words in the semantic lexicon are possible. Here, we report the results of two experiments where the α indexes of the words are used to filter expansions, with two different settings. In the first setting (Filter 1 hereafter) we compute α for each word of the original query, and we only add semantic neighbors for words of the original queries having α lower than a certain threshold. In the experiment, the threshold is fixed as the average α of the words of all the queries. In the second setting (Filter 2), we first filter words

	MAP	R-Prec	P@5	P@10	P@50	P@100
No expansion	21.78	30.93	92.80	89.40	79.60	70.48
with expansion	+13.80	+9.58	*+2.16*	+4.03	+5.58	+8.26
with expansion + Filter 1	+16.22	*+10.78*	*+3.02*	*+4.47*	*+9.20*	+12.51
with expansion + Filter 1 & 2	+22.83	+13.00	*+2.56*	+6.31	+14.10	+21.39

Table 4: Relative performance gain (%) on Tipster with query expansion with and without filtering; spectral lexicon

	MAP	R-Prec	P@5	P@10	P@50	P@100
No expansion	21.78	30.93	92.80	89.40	79.60	70.48
with expansion	+13.52	+9.50	*+2.59*	*+3.36*	+8.29	+9.99
with expansion + Filter 1	*+15.73*	*+9.27*	*+2.22*	*+4.96*	+9.63	+14.41
with expansion + Filter 2	+20.76	+13.63	*+3.88*	+5.82	+10.15	+14.27

Table 5: Relative performance gain (%) on Tipster with query expansion with and without filtering; Word2Vec

with Filter 1, and moreover, only semantic neighbors with α below a certain threshold are used to expand the queries.

Results with the spectral lexicon are given in Tab. 4, and those for WORD2VEC are in Tab. 5. Statistical significance (Wilcoxon with $p = 0.05$) are given: expansion results are compared with non-expanded version; expansion + filter 1 or 2 are compared with expansion (with no filtering). Non significant results are in italics.

The good results of expanding query (without filtering) with lexical resources are already known. Yet, it is worth noting that they are slightly made better, but not significantly, by filtering the words to be expanded (Filter 1) with the α index. When filtering also the words to add to the query (Filter 2), the gains are significantly better. These results holds for both lexical resources. In practice, a close examination of the expanded queries shows that words whose α are above the maximum threshold are indeed polysemic or general ones: *choice, term, use, way, young...*

6 Concluding remarks

In this article, we have shown how to adapt the notion of intrinsic dimensionality [14] to RSV similarities used in IR. In the follow up of [2], we have defined the α index to characterize indiscriminateness among neighbors of any point in the representation space. In a standard IR setting, we can exhibit the link between this index computed for any query and the performance of to be expected for this query. We have also applied this approach to word representation spaces, as generated by embedding techniques. We have shown how to improve query expansion by filtering the words to add to the query based on their indiscriminateness.

This work opens many research avenues. From a theoretical point of view, defining intrinsic dimensionality in the case of similarities used in IR, instead of distances studied by [14], raises questions. In this article, we have used the α index with the hypothesis that the RSV scores follow a Power Law distribution. Although this is experimentally verified, the precise link between intrinsic dimensionality and α should be formally investigated. From an applicative point of view, one can finds many use for this indiscriminateness notion. For instance, it may allow to propose LSI or LDA representation by adapted to the local complexity. Another pragmatic use would be to help formulate a query during an online search: when typing a word making the query α to raise, a user could be asked to precise or reformulate the query, which may improve the results as seen in Sect. 4.

References

1. Amati, G., Rijsbergen, C.J.V.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20, 357–389 (October 2002)
2. Amsaleg, L., Oussama, C., Furon, T., Girard, S., Houle, M.E., Kawarabayashi, K.I.: Estimating Local Intrinsic Dimensionality. In: 21st Conf. on Knowledge Discovery and Data Mining, KDD2015. Sidney, Australia (Aug 2015), https://hal.inria.fr/hal-01159217
3. Bellogín, A., de Vries, A.P.: Understanding similarity metrics in neighbour-based recommender systems. In: Proceedings of the 2013 Conference on the Theory of Information Retrieval. pp. 13:48–13:55. ICTIR '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2499178.2499186
4. Beygelzimer, A., Kakade, S., Langford, J.: Cover trees for nearest neighbors. In: Proc. of International Conference on Machine Learning (ICML). pp. 97–104 (2006)
5. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Review 51(4), 661–703 (2009)
6. Claveau, V., Kijak, E.: Direct vs. indirect evaluation of distributional thesauri. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 1837–1848. The COLING 2016 Organizing Committee, Osaka, Japan (December 2016), http://aclweb.org/anthology/C16-1173
7. Claveau, V., Kijak, E., Ferret, O.: Improving distributional thesauri by exploring the graph of neighbors. In: International Conference on Computational Linguistics, COLING 2014. Dublin, Irlande (Aug 2014), https://hal.archives-ouvertes.fr/hal-01027545
8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science (1990)
9. Fang, H., Tao, T., Zhai, C.: Diagnostic evaluation of information retrieval models. ACM Trans. Inf. Syst. 29(7) (2011)
10. Hersh, W., Buckley, C., Leone, T.J., Hickam, D.: Ohsumed: An interactive retrieval evaluation and new large test collection for research. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 192–201. SIGIR '94, Springer-Verlag New York, Inc., New York, NY, USA (1994), http://dl.acm.org/citation.cfm?id=188490.188557

11. Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for latent dirichlet allocation. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems 23, pp. 856–864. Curran Associates, Inc. (2010), http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf
12. Houle, M.E., Ma, X., Nett, M., Oria, V.: Dimensional testing for multi-step similarity search. In: Proc. of the 12th IEEE International Conference on Data Mining (ICDM). pp. 299–308 (2012)
13. Houle, M.E., Nett, M.: Rank cover trees for nearest neighbor search. In: International Conference on Similarity Search and Appli- cations (SISAP). pp. 16–29 (2013)
14. Houle, M., Kashima, H., Nett, M.: Generalized expansion dimension. In: Proc. of the 12th IEEE International Conference on Data Mining Workshops (ICDMW). p. 587–594 (2012)
15. Levina, E., Bickel, P.J.: Maximum likelihood estimation of intrinsic dimension. In: Advances in Neural Information Processing Systems (NIPS) (2004)
16. Lv, Y., Zhai, C.: Lower-bounding term frequency normalization. In: Proc. of the 20th ACM International Conference on Information and Knowledge Management. pp. 7–16. CIKM '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2063576.2063584
17. Metzler, D., Croft, W.: Combining the language model and inference network approaches to retrieval. Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval 40(5), 735–750 (2004)
18. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013). pp. 746–751. Atlanta, Georgia (2013)
19. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proc. of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '98). pp. 275–281 (1998)
20. Robertson, S.E., Walker, S., Hancock-Beaulieu, M.: Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In: Proc. of the 7^{th} Text Retrieval Conference, TREC-7. pp. 199–210 (1998)
21. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
22. Scholkopf, B., Smola, A.J., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5), 1299–1319 (1998)
23. Singhal, A.: Modern information retrieval: a brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (2001)
24. Strohman, T., Metzler, D., Turtle, H., Croft, W.: Indri: A language-model based search engine for complex queries (extended version). Tech. rep., CIIR (2005)
25. Turtle, H., Croft, W.: Evaluation of an inference network-based retrieval model. ACM Transactions on Information System 9(3), 187–222 (1991)
26. Venna, J., Kaski, S.: Local multidimensional scaling. Neural Networks (2006)
27. de Vries, T., Chawla, S., Houle, M.E.: Density-preserving projections for large-scale local anomaly detection. Knowledge Information Systems 32(1), 25–52 (2012)
28. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proc. of the SIGIR conference. pp. 334–342 (2001)