Information Retrieval manuscript No.
(will be inserted by the editor)

Reranking Web Search Results for Diversity

Ralf Krestel and Peter Fankhauser

Received: date / Accepted: date

Abstract Search engine results are often biased towards a certain aspect of a query or towards a certain meaning for ambiguous query terms. Diversification of search results offers a way to supply the user with a better balanced result set increasing the probability that a user finds at least one document suiting her information need. In this paper, we present a reranking approach based on minimizing variance of Web search results to improve topic coverage in the top-k results. We investigate two different document representations as the basis for reranking. Smoothed language models and topic models derived by Latent Dirichlet Allocation. To evaluate our approach we selected 240 queries from Wikipedia disambiguation pages. This provides us with ambiguous queries together with a community generated balanced representation of their (sub)topics. For these queries we crawled two major commercial search engines. In addition, we present a new evaluation strategy based on Kullback-Leibler divergence and Wikipedia. We evaluate this method using the TREC sub-topic evaluation on the one hand, and manually annotated query results on the other hand.

Our results show that minimizing variance in search results by reranking relevant pages significantly improves topic coverage in the top-k results with respect to Wikipedia, and gives a good overview of the overall search result. Moreover, latent topic models achieve competitive diversification with significantly less reranking. Finally, our evaluation reveals that our automatic evaluation strategy using Kullback-Leibler divergence correlates well with α-nDCG scores used in manual evaluation efforts.

Keywords Diversity, Web search, Reranking, Language models, Topic models, Variance, Diversity evaluation, Wikipedia

Ralf Krestel
L3S Research Center - Leibniz Universität Hannover
E-mail: krestel@L3S.de
Address: Appelstrasse 9a, 30167 Hannover, Germany

Peter Fankhauser
DFKI - German Research Center for Artificial Intelligence
Address: Stuhlsatzenhausweg 3, 66123 Saarbrcken, Germany

1 Introduction

Information Retrieval aims to provide the best possible results to meet a user's information need. Although keyword search and relevance rankings have been proven to be powerful tools to identify a user's interest and produce result lists with relevant pages, these mechanisms fail in certain situations. Ambiguous query terms are examples where relevant documents can not be reliably assessed without additional information from the user. Systems have to estimate a suitable relevance score and then rank accordingly. The most common way to produce these rankings is to follow the probability ranking principle (PRP) [22], which favors documents that are more likely to contain relevant information. For queries where the relevance scores for documents entail a lot of uncertainty, relevance rankings tend to leave a great deal of users unsatisfied: they abandon the query. Result diversification can reduce this effect [10] significantly.

Ambiguous queries are not the only reason why search engine results should reflect diversity. Queries like "Napoleon" or "immigration" are less ambiguous but rather multi-faceted. To capture different aspects of such queries a result set must contain diverse information and avoid semantically similar content within the top-k results. A truly diverse ranking then also offers an overview of the whole topic including various aspects and views.

Ideally, Web search results are not biased towards a certain interpretation or aspect. However, depending on the algorithm for assigning relevance scores certain interpretations of queries may be represented disproportionally high within the result set. For ambiguous queries such as "jaguar" or "java", commercial search engines nearly exclusively present documents about one interpretation ("car" and "programming"). Reducing the influence of the (manipulable) relevance score by combining it with a diversity aware component can help more users to find what they are looking for.

As described by Wang and Zhu [25], diverse rankings can be seen as the result of ranking under uncertainty where the user's information need cannot be ultimately defined. In the context of ambiguous queries, a system has to make a trade-off between the relevance of an isolated document and the risk involved of missing relevant aspects of a query. This task is tackled by Wang and Zhu by applying Modern Portfolio Theory [18], which is an economic theory that describes how to minimize the risk by not "putting all one's eggs in one basket", but on different investments. For ranking, this means to not favor one interpretation or aspect of a query over all others but prefer a *diverse* ranking.

Although diversifying Web search results has recently attracted a lot of interest within the research community [21, 25, 1, 13], automatic evaluation of diversity is still an open problem. Following [8], the TREC community has designed a task for subtopic retrieval in 2009 within the Web track [7]. The evaluation is based on subtopics of a query. These were identified using a query log of a commercial search engine and co-clicks, related queries and other information to find the different users' information need for each query. This also includes some manual judgement of the extracted subtopics. One of the drawbacks of this approach is the rather sparsely annotated data which makes it difficult to use for judging commercial search engines' results. The extraction process for the subtopics is also susceptible to missing aspects/subtopics of a query. The major drawback, however, is the need for manual judgement of whether a given Webpage covers a subtopic sufficiently or not. These judgements are cumbersome and costly.

In this paper we present a topic-centered approach for evaluation in contrast to the user-centered approach used in TREC. We propose an evaluation framework based on the Wikipedia encyclopedia and evaluate the diversity of Web search results for queries derived from titles of Wikipedia disambiguation pages. The coverage of the different aspects for a query present in Wikipedia is quantified using different entropy-based measures. We compare this evaluation setting with the TREC evaluation framework and a manually evaluation based on Wikipedia. We show that the obtained results are comparable with less costs and without having access to a large query log.

In addition, we present an approach to diversify search results by reranking. We estimate the relevance score of a document by its position in the original ranking and introduce a second score to reflect the additional diversity the document could add to the result list. This score is based on the variance of the underlying model for the document representation. We investigate language models and topic models [3], which have been shown to be useful document representations in the context of information retrieval tasks [27,26]. The main contribution of this paper are:

- We present an approach to reranking based on the original rank and the variance on two different document representations: Latent Dirichlet Allocation and smoothed language models.
- We propose and evaluate an evaluation framework to automatically assess diversity using ambiguous Wikipedia titles.
- We introduce entropy and Kullback-Leibler divergence as measures for diversity evaluation.

We show that the variance-based reranking outperforms the original rankings of two large commercial search engines with respect to diversity within the top-k results. Moreover, we show that latent topic models achieve competitive diversification requiring significantly less reranking. By comparing the proposed Wikipedia-based evaluation framework with the TREC subtopic retrieval evaluation we see comparable results without the need for a large-scale manual annotation effort.

2 Related Work

Search result diversification has received considerable attention in the past years; for recent overviews on the main issues and current approaches see, for example, [19].

2.1 Diversifying Search Results

One of the first works on result diversification introduces Maximum Marginal Relevance (MMR) as a ranking measure that balances relevance as the similarity between query and search results with diversity as the dissimilarity among search results [4]. Notably, MMR has not only been successfully used for diversity aware ranking, but also for text summarization, by selecting relevant and diverse text passages that cover the main topics or aspects of a text. Top-k diversification as pursued in this paper has the similar goal of covering the main aspects of a query by the top ranked search results.

Other approaches like [30] diversify recommendation lists to accommodate a users full spectrum of interests and minimize redundancy among the recommended items. Reranking approaches to diversify search results, e.g. [20] is based on query reformulations obtained from a query log where the focus lies on personalized search results, or [6] who describe a Bayesian reranking approach to maximize the coverage of different meanings of a query in the top 10 results have been explored before. Zhai and Lafferty [28] use statistical models for queries and documents. They model user preferences as loss functions and the retrieval process as a risk minimization problem. They retrieve models for subtopic retrieval that take dependencies between search results into account.

More recent approaches to diversification all essentially balance relevance with diversity, but differ in estimation of relevance and similarity, and choice of diversification objective. [1] classifies queries and results to categories of the ODP taxonomy, and diversifies results by maximizing the sum of categories covered by the top-k results weighted by the probability of categories given the query. Thereby the risk that the top-k results contain no relevant result for some category at all is minimized. [13] introduce a framework for analyzing approaches to diversification as variants of facility dispersion. On this basis they analyze and evaluate three diversification objectives: MaxSum, which takes into account all pairwise dissimilarities between top-k results as a measure for diversity, MaxMin, which maximizes just the minimum relevance and dissimilarity of results, and MonoObjective as a weighted aggregation of relevance and average dissimilarity for each top-k result. [25] introduces an approach to search result diversification adopting the Modern Portfolio Theory of finance. They generalize the well-known probability ranking principle (PRP) [22] by maximizing not only the relevance of top-k results but also minimizing the (co-)variance of the results. A greedy algorithm is used for ranking search results such that relevance is maximized while variance is minimized. [21] introduces a similar framework based on Portfolio Theory for reranking Web search results. Other than the greedy algorithm used in [25] they use quadratic programming optimization for arriving at optimal portfolios.

Very recently, [23] investigated the use of Wikipedia to improve diversity in Web search results. They manually annotated the top 100 results of a Web search engine for a set of 40 nouns with Wikipedia senses extracted from disambiguation pages. They showed that Wikipedia senses cover 56% of the Web pages and thus Wikipedia is much more suited than other sense inventories like WordNet (32%). Additionally, they propose using a vector space model and cosine similarity or word sense disambiguation algorithms to assign a Wikipedia sense to each page. Maximizing the number of different Wikipedia senses is then the goal of their greedy reranking algorithm.

There are a couple of approaches based on topically clustering the search results first and then diversifying based on the cluster information. Carterette and Chandar [5] propose to use probabilistic models to cover different facets of a query in the top-k results. Among others they use Latent Dirichlet Allocation (LDA) to cluster documents based on the extracted latent topics which they consider to be subtopics. They do not use a variance-based approach as proposed in this paper. Another clustering approach based on LDA is presented in [15]. For each query LDA is applied and documents are assigned to latent topics with each topic constituting a cluster. For the diverse ranking, clusters are ranked, from which documents are picked. A cluster approach based on k-means clustering is described in [2]. The documents are clustered and the diversified ranking is produced by picking documents from each cluster based on its size.

The problem of result diversification is also investigated in the area of structured data queries. Recommending a set of items to the user or returning a list of products in response to a keyword query are applications for result diversification. Vee et. al. [24] propose an efficient algorithm to find a representative, diverse set of top-k results for a given form-based query. All attributes of an object are ordered according to their priority for diversification by a domain expert. Jain et. al. [16] make use of k-nearest neighbor clustering techniques and combine it with a notion of diversity based on a distance metric. Each query is represented as a point in an n-dimensional space and the k-nearest neighbors are selected which also satisfy the required distance. Demidova et. al. [11] go a step further by introducing an approach that diversifies keyword queries against structured databases based on their schema rather than diversifying the results. A necessary condition for these approaches is that the database schema captures the semantics of the domain at hand.

In this paper we follow the approach of Wang and Zhu [25] to minimize (co-)variance for maximizing diversity, but rely on the search engine ranking for estimating relevance rather than estimating it from the documents directly. As the relevance estimated from the original ranking and the variance are typically on a different scale, this requires to normalize the variance, in order to balance relevance and variance. Moreover, we evaluate to what extent a condensed representation in terms of latent topic models can capture diversity better than the language modeling approach used in [25].

2.2 Diversity Evaluation

Evaluation of result diversification requires new measures that consider more than just simple relevance judgements. To this end, several extensions to traditional measures have been proposed. Their common idea is that queries and documents cover several subtopics (also called aspects or nuggets), and thus relevance is assessed w.r.t. subtopics rather than w.r.t. documents.

[6] evaluate their approach on different TREC tasks (robust track, interactive track, and manually annotated TREC data). In [20] user assessment of the result is used to measure whether the diversified result list contains at least one document satisfying the user's interest. [29] introduce variants of recall and precision that take into account the subtopics of a query. S-recall at K measures the proportion of subtopics covered by the top-k results, and S-precision at r measures the ratio of the best rank K_{opt} that can be achieved for a given recall r and the actual rank K with recall r.

[8] introduce α–nDCG as a generalization of the nDCG measure (normalized Discounted Cumulative Gain). Whereas nDCG only measures the relevance of search results, discounted by the logarithm of their rank, α–nDCG in addition penalizes repeated subtopics in search results. For evaluating diversification of search engine results the required explicit relevance assessments in terms of subtopics are difficult to acquire. [13] avoid the need for human relevance assessment by taking Wikipedia pages returned for a query as subtopics, and estimate subtopic relevance by a thresholded similarity between result documents and Wikipedia pages to measure S-recall (also called novelty). In this paper we also compare original and diversified rankings with respect to Wikipedia, but estimate "subtopic coverage" directly on the language models of the top-k results and Wikipedia.

3 Diversification by Reranking

Our goal is on the one hand to cover for each query as many *different aspects* as possible within the top-k search results. On the other hand, ranking of Web pages is predominantly done by picking the most *topically relevant*[1] pages for a keyword query according to the probability ranking principle [22]. A diverse search result cannot neglect the relevance aspect. Thus, the relevance of Web pages for a user's query still plays an important role. A trade-off between relevance and diversity [9] is incorporated within our system to accommodate this mutual relation.

[1] Possibly, commercial search engines also include popularity and other factors in their ranking.

3.1 Overall Approach

In its most simple form the probabilistic ranking principle assumes that the usefulness of each individual result only depends on the query, but does not depend on the other results. Under this assumption, given a good estimate of the relevance $E(r_i)$ for each result r_i individually, ordering the results by decreasing $E(r_i)$ is optimal. However, especially for Web search results, this assumption clearly does not hold. In the extreme, if the most relevant result is duplicated, the top results will all be the same, with all but the first one not adding useful information. More generally, if results overlap with each other, the top results will often be pre-occupied by one interpretation of a query. Thus, the general goal of diversification is to balance between relevance of individual results and their overlap.

One popular approach to this end is to minimize the mutual overlap between the top k results, using some similarity measure such as Jaquard similarity or cosine similarity. We adopt a closely related approach, originally introduced in [25], which *maximizes the expected relevance* $E(R_k)$ and *minimizes the variance* $Var(R_k)$ for the top k documents of a search result $R_n = r_1, ..., r_n$:

$$E(R_k) - B * Var(R_k) \tag{1}$$

where B regulates the trade-off between relevance and diversity. Expected relevance $E(R_k)$ and variance $Var(R_k)$ are calculated as weighted sum over the individual results r_i:

$$\begin{aligned} E(R_k) \quad &= \textstyle\sum_{i=1}^{k} w_i E(r_i) \\ Var(R_k) &= \textstyle\sum_{i=1}^{k} \sum_{j=1}^{k} w_i w_j c_{i,j} \end{aligned}$$

where $c_{i,j}$ is the covariance of results r_i, r_j, which is calculated based on their vector representation (see Equation 7), and w_i is a normalized discount factor [17]:

$$w_i = \frac{1}{log_2(i+1) \sum_{j=1}^{n} \frac{1}{log_2(j+1)}} \tag{2}$$

$\frac{1}{log_2(i+1)}$ is 1 for rank $i = 1$ and decreases monotonically, the second factor in the denominator normalizes the sum of all w_i to 1.

Diversity is inversely proportional to variance: A small variance $Var(R_k)$ corresponds to large diversity, because all diverse aspects of a query are covered more or less equally. In the extreme, when all aspects, as represented by their topical terms (see Section 3.2.1) occur equally often, the variance is 0. B controls the relative importance of diversity vs. relevance. For $B > 0$ relevance and diversity are balanced against each other. In particular for ambiguous queries, choosing relevant and at the same time diverse and complementary documents with high $E(R_k)$ and low $Var(R_k)$ reduces the risk that the top k results do not contain any relevant document at all for some of the possible query interpretations. With very large B, the original ranking is practically overrun, and the top (few) k results will cover any topic that occurs somewhere in the complete search result. However, in our experiments giving equal weight to the original ranking and variance typically achieves good topic coverage, which does not significantly improve with increasing B (see Section 5.2. On the contrary, large B can even hurt topic coverage, because documents with low relevance very often do not cover any relevant topic at all. For $B < 0$ relevance and variance are maximized, and thus diversity is minimized. This favors one particular interpretation with high $E(R_k)$ but also high $Var(R_k)$, which increases the risk of missing out other plausible interpretations altogether.

Finding a reranking that globally optimizes the objective in Equation 1 is infeasible, as it would require testing all permutations of the original ranking. Thus, following common practice, we approximate the optimal reranking using a greedy algorithm that selects for each new rank k the result r_i such that the increase in the objective at rank k ($O_k - O_{k-1}$) is maximized [25]:

$$\begin{aligned} O_k - O_{k-1} = & \sum_{i=1}^{k} w_i E(r_i) - B \sum_{i=1}^{k} \sum_{j=1}^{k} w_i w_j c_{i,j} \\ & - \sum_{i=1}^{k-1} w_i E(r_i) + B \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} w_i w_j c_{i,j} \\ = & \ w_k (E(r_k) - B w_k c_{k,k} - 2B \sum_{i=1}^{k-1} w_i c_{i,k}) \\ \propto & \ E(r_k) - B w_k c_{k,k} - 2B \sum_{i=1}^{k-1} w_i c_{i,k} \end{aligned} \quad (3)$$

The multiplier w_k is constant for all candidate documents to be selected for rank k and thus can be ignored.

In contrast to [25] we do not estimate the expected relevance $E(r_k)$ from the query and individual results, but rather rely on the original ranking of the search engine, which takes into account a variety of factors, including relevance, popularity, and user preferences. As search engines typically do not provide an actual score we set $E(r_k)$ to the discount factor w_i of a result document r_i to be reranked to position k. $c_{k,k}$ is the (inner) variance $\sigma^2(r_k)$ of result r_k at the new rank k, as defined in Equation 7. This leads to the following optimization objective: At each new rank k select the document r_i at the original rank i, such that

$$w_i - B w_k \sigma^2(r_i) - 2B \sum_{j=1}^{k-1} w_j c_{j,i} \quad (4)$$

is maximized.

A couple of technical statements are in order: To effectively balance $E(R_k)$ and $B * Var(R_k)$ they should be in the same order of magnitude. To this end, we calibrate B as follows:

$$B = \frac{\beta}{avg_i \sigma^2(r_i)} \quad (5)$$

where $avg_i \sigma^2(r_i)$ is the average (inner) variance over all results r_i. With this approach, $\beta = 1$ gives approximately equal weight to relevance and diversity[2].

The complexity of the greedy reranking algorithm is $O((n-k) * k * |V|)$ for reranking in the top-k results, given n overall results and vocabulary size $|V|$. Thus for relatively small k in the range of the typical 10 results on the first page online reranking is feasible, in particular, when combined with standard techniques such as caching popular queries.

[2] With $\beta = 1$, $\sum_{i=1}^{n} O_i \approx \sum_{i=1}^{n} w_i - \frac{\beta \sum_{i=1}^{n} w_i \sigma^2(r_i)}{avg_i \sigma^2(r_i)} = 0$. This assumes that the overall sum of covariances is zero, which is probably an underestimation.

3.2 Representation of Documents

In order to calculate the variances we represent individual documents r_i as vectors. We have experimented with two alternative representations: Smoothed (unigram) language models and latent topic models.

3.2.1 Language Models

The Jelinek-Mercer smoothed language models [27] for a document r are defined as

$$q_i = \lambda * p(v_i|r) + (1 - \lambda) * p(v_i) \tag{6}$$

where $p(v_i|r)$ is the relative frequency of term v_i in r, and $p(v_i)$ is the relative collection frequency of v_i. For smoothing we use the relatively large[3] $\lambda = 0.99$.

Given two vectors $U = u_1 \dots u_n$ and $Q = q_1 \dots q_n$, their co-variance is defined as:

$$\begin{aligned} Var(U, Q) &= \frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})(q_i - \bar{q}) \\ &= \frac{1}{n} \sum_{i=1}^{n} u_i q_i - \frac{1}{n^2} \end{aligned} \tag{7}$$

The simplification is based on $\bar{u} = \bar{q} = 1/n$.

It is interesting to compare this to cosine similarity used by other approaches to diversification:

$$Cos(U, Q) = \frac{\sum_{i=1}^{n} u_i q_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \sqrt{\sum_{i=1}^{n} q_i^2}}$$

As can be seen, covariance and cosine similarity differ only w.r.t. normalization, which plays a minor role when operating on vectors representing a normalized probability distribution. However, whereas minimizing the mutual cosine similarity between results only accounts for the overlap between results, minimizing the overall variance of a result list also accounts for the inner variance of individual results. Thereby, results that cover more aspects of a query will tend to be ranked higher.

3.2.2 Latent Dirichlet Allocation

Smoothed language models may suffer from the curse of dimensionality, and thus not properly represent the topics or aspects of a result list. As a consequence, variance measured directly on the bag of words may not be a good indicator for topical coverage. For example, if two results are about the same topic, but use different vocabulary, their covariance will be underestimated.

Thus as an alternative representation, we have also experimented with Latent Dirichlet Allocation (LDA) [3], which maps documents to a mixture of only a few latent topics. Variance is then estimated on the much lower dimensional representation of the latent topics

[3] Since we need smoothing only for avoiding zero probabilities in our evaluation based on Kullback-Leibler Divergence, we have chosen an unusually large λ. Smaller λs would just make the individual documents more similar, and thereby reduce their (co-)variance.

$P(z_i = j \mid d_i)$ as defined in Equation 11 rather than on the bag of words derived from Equation 6.

The principal idea behind LDA is based on the hypothesis that a person writing a document has certain topics in mind. To write about a topic then means to pick a word with a certain probability from the pool of words of that topic. A whole document can then be represented as a mixture of different topics. For Web pages where the author of a document can be considered one entity, these topics reflect the entity's view of this document and her particular vocabulary.

The modeling process of LDA can be described as finding a mixture of topics for each Web page, i.e., $P(z \mid d)$, with each topic described by terms following another probability distribution, i.e., $P(t \mid z)$. This can be formalized as

$$P(t_i \mid d) = \sum_{j=1}^{Z} P(t_i \mid z_i = j) P(z_i = j \mid d), \tag{8}$$

where $P(t_i \mid d)$ is the probability of the ith term for a given document d and z_i is the latent topic. $P(t_i \mid z_i = j)$ is the probability of t_i within topic j. $P(z_i = j \mid d)$ is the probability of picking a term from topic j in the document. The number of latent topics Z has to be defined in advance and allows to adjust the degree of specialization of the latent topics. LDA estimates the topic-term distribution $P(t \mid z)$ and the document-topic distribution $P(z \mid d)$ from an unlabeled corpus of documents[4] using Dirichlet priors for the distributions and a fixed number of topics. Gibbs sampling [14] is one possible approach to this end: It iterates multiple times over each term t_i in document d_i, and samples a new topic j for the term based on the probability $P(z_i = j | t_i, d_i, z_{-i})$ based on Equation 9, until the LDA model parameters converge.

$$P(z_i = j \mid t_i, d_i, z_{-i}) \propto \frac{C^{TZ}_{t_i j} + \beta}{\sum_t C^{TZ}_{tj} + T\beta} \frac{C^{DZ}_{d_i j} + \alpha}{\sum_z C^{DZ}_{d_i z} + Z\alpha} \tag{9}$$

C^{TZ} maintains a count of all topic–term assignments, C^{DZ} counts the document–topic assignments, z_{-i} represents all topic–term and document–topic assignments except the current assignment z_i for term t_i, and α and β are the (symmetric) hyperparameters for the Dirichlet priors, serving as smoothing parameters for the counts. Based on the counts the posterior probabilities in Equation 8 can be estimated as follows:

$$P(t_i \mid z_i = j) = \frac{C^{TZ}_{t_i j} + \beta}{\sum_t C^{TZ}_{tj} + T\beta} \tag{10}$$

$$P(z_i = j \mid d_i) = \frac{C^{DZ}_{d_i j} + \alpha}{\sum_z C^{DZ}_{d_i z} + Z\alpha} \tag{11}$$

In our evaluation we experimented with different numbers of topics, and achieved best results with 1000 topics for the entire search result, from which only few topics were associated to each individual result.

Table 1 Top 10 search results for query “Caesar” using Google search engine. Note that topic Julius Caesar also covers a variety of sub-topics.

Rank	Original	LDA	LM
1	Caesars Palace Hotel	Julius Caesar Biography	A Weblog by Julius Caesar
2	Caesars Palace Hotel Shopping	Caesar III Heaven Games	Julius Caesar Biography
3	Caesars Windsor Hotel	Free Website on Community Architect	Caesar’s Campaigns in Gaul
4	Gaius Julius Caesar Biography	Shakespear’s Julius Caesar	Commentariorvm de bello gallico
5	COADE CAESAR II - Pipe Stress Analysis	A Weblog by Julius Caesar	Shaw’s Caesar and Cleopatra
6	COADE Company	COADE CAESAR II - Pipe Stress Analysis	Augustus Biography
7	Free Website on Community Architect	Littel Caesar Movie 1930	CAESAR Anthropometry
8	Julius Caesar: Guide to Online Resources	Commentariorvm de bello gallico	Littel Caesar Movie 1930
9	Caesar Augustus: Guide to Online Resources	Caesar’s Campaigns in Gaul	Julius Caesar Biography
10	Caesar Miniaturs Company	Shakespear’s Julius Caesar Paraphrase	Svetoni tranqvilii vita divi ivli

4 Evaluating Diversity

To evaluate our approach we propose to use Wikipedia as a source of ground truth for diversity. Wikipedia has been shown to be an effective and reliable source of semantic knowledge [12] and was used before in the context of diversity evaluation [13]. We think that this kind of evaluation is superior to manually selected corpora to judge the diversity of Web search result rankings. Hand-crafted collections like the one used for the TREC subtopic retrieval task are not as complete and representative as a community maintained encyclopedia like Wikipedia.

For the evaluation, we compare the original ranking with the diversity-oriented reranking and a baseline reranking based on a simple notion of relevance. The test queries are taken from the titles of Wikipedia disambiguation pages. The basic assumption is that Wikipedia articles cover the major alternative interpretations of ambiguous queries. This claim was recently backed by [23], who showed that more than 50% of pages in their test set can be assigned to Wikipedia pages representing a particular sense of the query. Moreover, we also compare the various rankings with the “complete” search result returned for each query.

We conducted several experiments to evaluate our reranking algorithm and to verify our evaluation approach:

1. Reranking based on language models of search results.
2. Reranking based on topic models derived from Latent Dirichlet Allocation.
3. Comparing diversity of result rankings from Google and Yahoo!.
4. Comparing our evaluation using TREC data and manual judgement.

4.1 Testdata

To evaluate diversity we are interested in queries that have a broad variety of aspects. This does not neccessarily mean that the queries are ambiguous. A keyword query like “Las Vegas” might have different meanings but even the interpretation as the name of a city has a lot of aspects and subtopics which diversity aware search engines should cover in the top-k results.

The generation of the ground truth testdata was a two phase process. Firstly, we took the Wikipedia disambiguation pages and removed all pages containing digits in the title (e.g. Wikipedia page “442_(disambiguation)”). Secondly, we searched in a Wikipedia MYSQL-dump with the title of the disambiguation page in the title field of the database. All titles returning between 10 and 100 Wikipedia pages were kept and the others discarded. We sorted the titles of the disambiguation pages by the sum of the inlink degree of the Wikipedia

[4] In our case, the corpus for each query consists of the top 700 result pages returned by the search engine

Table 2 Wikipedia pages containing the query “Billboard” and its corresponding link indegrees.

Titles of Wikipedia pages containing “billboard”	Link Indegree
Billboard magazine	2100
Billboard Hot 100	932
Billboard 200	323
Billboard (advertising)	74
Billboard Music Award	24
Adult Contemporary (Billboard Chart)	16
Billboarding	2
Billboard Liberation Front	2
Billboard antenna	2
Billboard toppling	1
List of most frequently mentioned brands in the Billboard Top 20	1
Billboard Utilising Graffitists Against Unhealthy Promotions	0
Billboard Comprehensive Albums	0
Billboards of Lahore	0

pages. The top 240 titles constitute our query set and the corresponding Wikipedia pages are our ground truth data. One example query (“billboard”) with its corresponding Wikipedia page titles is shown in Table 2.

To get the ranking of the commercial search engines from Google and Yahoo! we crawled[5] the result lists for each query up to rank 1000. For the Yahoo! search engine we got an average of 628 results per query; for Google we got 730. Search results from Wikipedia were discarded, in order to assess original and diversified rankings of non-Wikipedia results. Also all pages without textual content were removed from the collection. In addition, we removed boilerplate text from the result Web pages using boilerpipe[6], an open source library for extracting fulltext from HTML pages, to obtain clean content for each page. For both search engines we got the original rankings for each query ordered by rank. For Wikipedia as well as for the search engine results we removed stopwords[7]. For each query we also computed a reranking of the original results based on a simple relevance assessment taking the term frequency for each query term into account.

For each rank k we define R_k as the concatenation of all documents $r_i, 1 \leq i \leq k$, after removal of Wikipedia results and results without textual content. The smoothed language model Q for each R_k is computed as described in Equation 6.

The language models U for Wikipedia are in addition weighted by the logarithm of the indegree of articles (d_j), in order to push more prominent interpretations [8]

$$p(v_i|W) = \frac{\sum_{j=1}^{m} log_2(d_j) * n(v_i, w_j)}{\sum_{j=1}^{m} log_2(d_j) * |w_j|} \qquad (12)$$

where m is the number of Wikipedia result pages for a query, $n(v_i, w_j)$ is the frequency of word v_i in article w_j, and W signifies that the language model is conditioned on Wikipedia.

[5] For Google we crawled the page manually; for Yahoo! we used the API; both were crawled in January 2010

[6] `http://code.google.com/p/boilerpipe/`

[7] The stopword list is an extended version of the list available at `http://truereader.com/manuals/onix/stopwords1.html` and contains about 600 entries

[8] This follows the observation in [23] that the indegree in Wikipedia correlates with the overall frequency of an interpretation.

Unless otherwise noted, Q refers to the language model of top-k search results, S refers to the complete search result of a particular query, and U refers to Wikipedia articles which contain the query in their title.

4.2 Evaluation Measures

As a measure for how well the top-k Web search results for a query approximate the corresponding[9] Wikipedia articles we calculate the Kullback-Leibler divergence between the smoothed unigram language models for the top-k results and for Wikipedia articles. This measure estimates the number of additional bits needed to encode the distribution $U = u_1 \dots u_n$, using an optimal code for $Q = q_1 \dots q_n$, where $n = |V|$ is the combined vocabulary size.

$$\begin{aligned} D_{KL}(U||Q) &= H(U;Q) - H(U) \\ &= \textstyle\sum_{i=1}^{n} u_i * log_2(\frac{u_i}{q_i}) \end{aligned} \tag{13}$$

In our setting, distribution Q is the combined language model of the top-k search results and distribution U is the language model for the Wikipedia articles. Thus $D_{KL}(U||Q)$ can be directly used to measure the similarity with the combined Wikipedia articles and assess the coverage of the top-k Web pages with respect to Wikipedia.

To assess the effect of diversification on the search results Q, we also measure the entropy $H(Q)$ for the different rankings. The higher the entropy of the top-k results, the more diverse is the set of top-k Web pages.

$$H(Q) = -\sum_{i=1}^{|V|} q_i * log_2(q_i) \tag{14}$$

Spearman's rank correlation coefficient ρ is used to quantify the degree of reranking between two rankings x and y.

$$\rho(x,y) = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)} \tag{15}$$

where x_i and y_i are the ranks at position i, and n is the number of results. A value of 1.0 means perfect correlation, 0.0 no correlation and -1.0 perfect negative correlation. In our setting, we are interested in the degree of reranking performed by the different algorithms with respect to the original ranking.

4.3 An Example for Diversification

To exemplify the effect of diversification, we randomly picked the query "Caesar" from our evaluation set. Table 1 gives the top 10 results for this query by the original ranking and by rankings diversified on the basis of latent topic models and language models[10]. The different colors reflect a broad categorization of the pages. While the original ranking covers

[9] weighted combination of language models of Wikipedia pages returned by the Wikipedia search engine for a query

[10] In all rankings the Wikipedia entries have been removed.

some aspects of this query, including the historical persons “Julius Caesar” and “Caesar Augustus”, hotels named “Caesar”, and other companies using the iconic label “Caesar”, both diversified rerankings arguably cover also other aspects, including movies and dramas about “Caesar”, pointers to Julius Caesar’s literary work, and also a broader variety of companies labeled “Caesar”, with the notable exception of hotels. For other queries we can observe a similar effect. Generally, the diversified reranking achieves a better topic coverage in the top 10 results compared with the original ranking.

5 Results

We thoroughly analyzed the results and were particularly interested in three aspects: *(1)* Comparing diversity using language models and topic models described in Section 5.1, *(2)* balancing relevance and diversity (Section 5.2), and *(3)* comparing the diversity of Google and Yahoo! (Section 5.3).

5.1 Diversification by LM vs. LDA

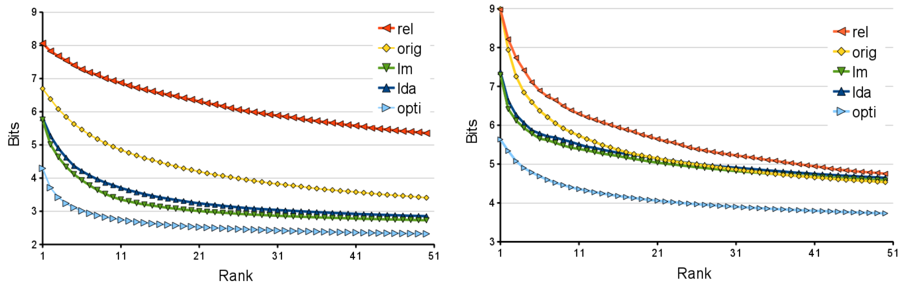
The goal of our first evaluation is to assess the effect of diversification for the two proposed models. Figures 1(a) and 1(b) show the Kullback-Leibler divergences $D_{KL}(U||Q)$ and $D_{KL}(Q||U)$ between the aggregated Wikipedia language models U and various rankings Q for ranks $k = 1..51$, averaged over all 240 queries in our testset. The original ranking from Google is labeled $orig$. For the “optimal” ranking $opti$, we greedily reranked search results such that $D_{KL}(U||Q)$ is minimal for each rank k. For reranking we used $\beta = 1$, balancing relevance and diversity evenly.

As is to be expected, the ranking rel based solely on relevance has the largest divergence D_{KL} to Wikipedia in both directions. Focussing only on relevance while disregarding possible redundancies between individual results leads to a bad topical coverage in the first few results. The original ranking $orig$ has the second largest divergence, and the optimal ranking $opti$ has the smallest divergence. The diversified reranking using language models lm slightly outperforms latent topic models lda at all ranks. However, this comes at the cost of a significantly larger amount of reranking: The average Spearman’s rank correlation coefficient ρ between lda and $orig$ is 0.23, which is more than twice of $\rho = 0.09$ between lm and $orig$[11] Interestingly, also the “optimal” reranking $opti$ has a significantly higher $\rho = 0.17$.

Figure 2(a) shows how quickly the various rankings Q approximate the language model of the overall search result S for each query. The smaller the divergence for the top k results the better they represent the overall result. Again, the relevance based ranking rel shows the highest divergence overall, followed by the original ranking $orig$. But the optimal ranking $opti$ is surpassed by lm at rank 12, and by lda at about rank 25. Thus, optimizing w.r.t. Wikipedia content of a query generally also achieves a better representation of the search result in the first few ranks, but the generic diversification by minimizing variance performs slightly better for higher ranks (The plot for $D_{KL}(Q||S)$ is very similar).

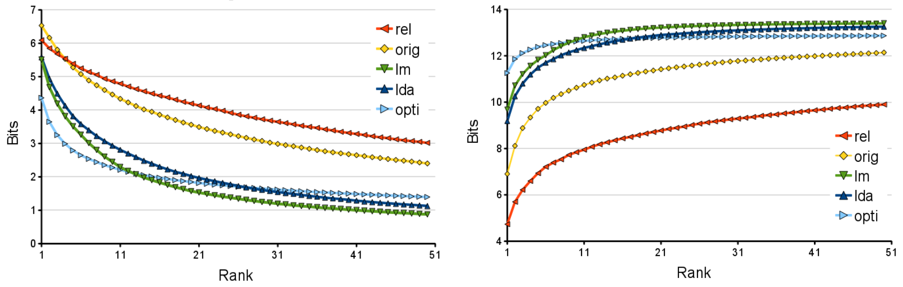
The Kullback-Leibler divergence only measures the additional bits needed for representing query result distribution Q given an optimal code for the Wikipedia distribution U, i.e., it explicitly disregards the entropy $H(Q)$. Figure 2(b) shows the entropy for the various rerankings. As is to be expected, reranking by minimizing the variance leads to a higher

[11] The difference is significant (Conf. of 0.99 based on a paired t-test on the correlation coefficients for the 240 queries).



(a) $D_{KL}(U||Q)$ (b) $D_{KL}(Q||U)$

Fig. 1 Kullback-Leibler divergence between the top-k search results (Q) and Wikipedia (U)



(a) $D_{KL}(S||Q)$ (b) $H(Q)$

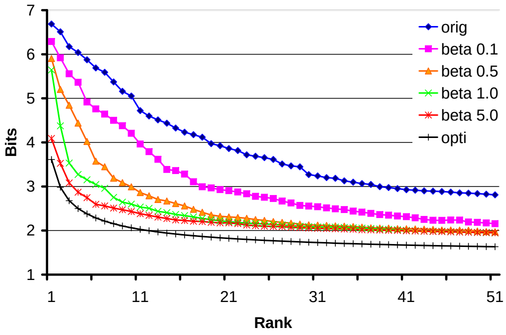
Fig. 2 Kullback-Leibler divergence between the top-k search results and the complete (600 pages per query) search results (left) and entropy of the top-k search results (right)

entropy $H(Q)$ at all ranks. Also the optimal ranking $opti$ leads to a higher entropy, but levels out at a slightly lower entropy than for the diversified ranks. Naturally, the increased entropy $H(Q)$ also leads to an increased cross-entropy $H(Q;U)$ (not shown). One consequence of this is that the improvement in divergence by diversification is less pronounced for $D_{KL}(Q||U)$ (see Figure 1(b)) than for $D_{KL}(U||Q)$. After about rank 15, the gain of diversification is balanced by the cost of diversification in terms of entropy. The entropy of the ranking rel based on releance is by far the smallest at all ranks. Even at rank 50 it just reaches the entropy of the diversified rerankings at rank 1. This again illustrates that ranking based on relevance only covers only few aspects of the search result.

The effects of diversification for Yahoo! search results are similar; see Section 5.3 for a comparison of Yahoo! and Google.

5.2 Balancing Relevance and Diversity

In this section we analyse the effect of the parameter β, which balances between relevance and diversity. To this end, we selected 10 queries, where the difference between divergence of the top 10 results and of the complete result is maximal and varied β between 0.1 and 5. Figure 3 (left) compares the divergence using language models as document representations and shows how the KL-Divergence of the rerankings with different βs lie in between the



β	**LDA**	**LM**
0.1	0.56	0.29
0.5	0.33	0.10
1.0	0.27	0.07
5.0	0.16	0.03
Opt	0.24	0.24

Fig. 3 Kullback-Leibler divergence $D_{KL}(U||Q)$ for different β using language model reranking (left) and Spearman's rank correlation coefficient for different β and for the optimal ranking

original ranking and the optimal ranking. Increasing β beyond 5.0 does not further improve the results.

The right table in Figure 3 compares the rank correlation for both search engines and for the optimal ranking. The general behaviour is consistent. The divergence decreases at all ranks with increasing β at the cost of a higher degree of reranking, resulting in a lower rank correlation ρ. $\beta > 1$ achieves only a relatively small improvement, $\beta > 5$ (not shown) achieves no further visible improvement. As already observed in Section 5.1, diversification based on latent topic models lda generally achieves a ranking closer to the original ranking than diversification based on language models lm.

5.3 Comparing Two Search Engines

Search engines certainly also make an effort towards covering the most important aspects of queries as one of their optimization objectives. Our evaluation framework can also be used to compare topic coverage in the top-k results for different search engines. Figure 4 shows the difference $D_{KL}(\text{Google}) - D_{KL}(\text{Yahoo!})$ of the two evaluated search engines of the symmetric Kullback-Leibler divergence:

$$D_{KL}(U,Q) = \frac{D_{KL}(U||Q) + D_{KL}(Q||U)}{2} \tag{16}$$

One graph shows the divergence with respect to Wikipedia and the other one with respect to the complete search result. Apparently Google search results tend to be significantly less diverse than Yahoo!'s search results; in the top ranking positions the divergence of Google is almost 1 bit higher than the divergence of Yahoo!.

Of course such a comparison should not be taken as evidence on any inherent bias of a search engine. Firstly, the observable difference may in part be due to different strategies of including Wikipedia pages, which were discarded for evaluation. In particular, if a search engine tends to rank Wikipedia pages on top, and diversifies the next few results w.r.t. the top results, discarding Wikipedia pages from the evaluation will lead to understimating the topical coverage of the remaining results. Secondly, the two search engines employ slightly different strategies in grouping related search results, which were not taken into account in our evaluation, where we mapped search results to a flat ranked list. Finally, of course Wikipedia does not necessarily cover all possible interpretations of a query.

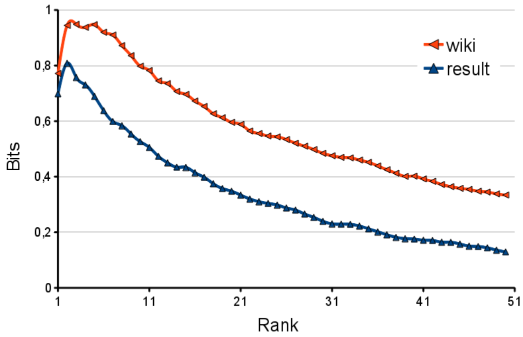


Fig. 4 Comparison of the diversity of Google and Yahoo! using symmetric $D_{KL}(\text{Google}) - D_{KL}(\text{Yahoo!})$

6 Evaluating Wikipedia-Based Evaluation

To verify the viability of our proposed diversity evaluation based on Kullback-Leibler divergence and Wikipedia, we compare the results with two other diversity evaluation frameworks: Subtopic retrieval from TREC and a manual evaluation using hand annotated search results from Santamaría et al. [23].

6.1 Comparison with TREC Evaluation

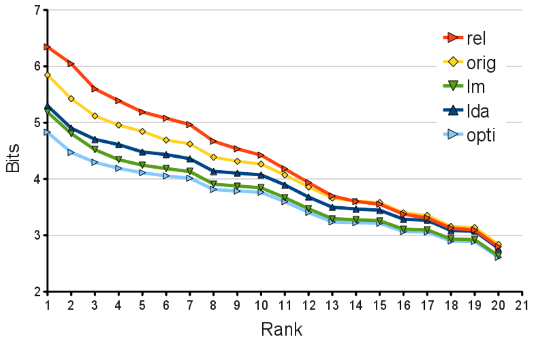
In order to assess our proposed evaluation criterion based on Wikipedia coverage, we have applied our diversification approach on TREC data. In the Web Track 2009, TREC introduced a dataset to evaluate subtopic coverage of rankings [7]. They provided a Web crawl, 50 queries, and automatically extracted subtopics for these queries. This extraction was done using the query log of a commercial search engine, co-click data, and other information. A set of Webpages from the crawl was then annotated manually with the relevant subtopic or with "not relevant" in case the page does not cover any subtopic.

To compare this evaluation framework with our Wikipedia-based approach we identified a subset of the data satisfying our requirements:

1. A query using Wikipedia's search mechanism must return at least 100 Wikipedia pages.
2. An annotated document must occur in the top 1000 results of Google.

Among the 50 TREC queries, 7 did not yield any Wikipedia page, 8 less then 10, and 8 less then 100 result pages when searching Wikipedia. This leaves us with 27 queries with up to 500 ranked, relevant Wikipedia pages. The average overlap of Web search results from our crawl with annotated TREC pages matched by URL is 26.6 pages per query for Google. This leaves us with only a few pages annotated as relevant for a specific subtopic and many queries with no annotated pages for a certain subtopic.

Figure 5 (left) shows how the original ranking, and the introduced reranking approaches approximate the corresponding Wikipedia content for the 27 TREC queries (c.f. Figure 1(a)). Again, by construction, the optimal ranking covers Wikipedia content best in the top k results. However, for the TREC queries, diversification based on the LDA topics slightly outperforms diversification based on the language models. The original ranking depicts the largest distance to Wikipedia.



	@ 5	@ 10	@ 20
Rel	0.198	0.229	0.260
Orig	0.214	0.254	0.306
LM	0.205	0.255	0.299
LDA	0.216	0.253	0.298
Opt	0.254	0.290*	0.329*

Fig. 5 Kullback-Leibler divergence $D_{KL}(U||Q)$ (left) and α–nDCG values (right) for the TREC evaluation

Using the manually assessed subtopics and evaluation scripts provided by the TREC organizers, we computed α–nDCG [8], shown in Table 5 (right). On this small dataset the relevance based ranking clearly performs worst, while the original ranking and the rerankings by means of minimizing variance perform rather similarly. The slight differences are not statistically significant. Only the "optimal" reranking achieves a significant improvement for all metrics but α–nDCG@5 according to a 2-tailed paired t-test with confidence well above 95% (marked with asterisks in the table). This indicates that diversification based on a more or less representative goal model, such as Wikipedia, can outperform diversification based on analyzing only the search results. Investigating and evaluating such a goal-driven approach to diversification in more detail is an interesting subject for future work.

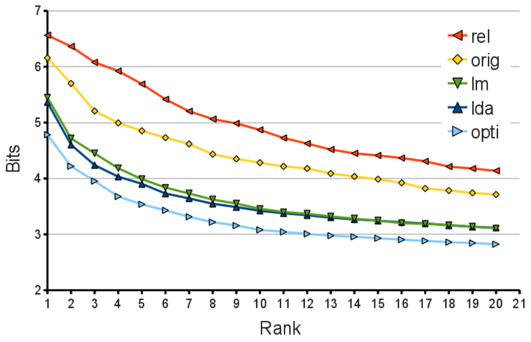
In summary, for the subset of the TREC queries where we had enough data, diversification based on latent topic models generally achieves better coverage of Wikipedia than diversification based on language models. Probably due to the rather small overlap between manual TREC assessments and the search engine results, the original rankings and rerankings achieved similar performance with respect to α-nDCG.

To put this into perspective, we note that the clearly best run in the diversity task of TREC 2009 [7] also just took the original ranking provided by a major commercial search engine. Thus the achieved improvement over the original ranking is fairly remarkable.

6.2 Comparison with Manual Evaluation

As a second dataset to validate our evaluation method we used a test corpus compiled by Santamaría et al. [23]. This corpus comprises Web search results for 40 ambiguous queries consisting of 15 ambiguous nouns from the Senseval-3 dataset and 25 additional ambiguous nouns, where one of the senses is a band name. For all senses there exists a corresponding Wikipedia article. For each query the top 150 documents have been manually annotated with one or more senses. Documents with little text, disamgiguation pages, and documents not corresponding to any Wikipedia sense have been discarded.

On the basis of the manual annotations, we have again computed α–nDCG. Figure 6 (right) compares the averaged α–nDCG for the 40 queries with our proposed evaluation criterion of Wikipedia coverage measured by the Kullback-Leibler Distance between the search result and the language model of Wikipedia articles (Figure 6 (left)). As can be seen, the relative performance of the various rerankings is the same for both evaluation measures, in particular at smaller ranks. The original ranking $orig$ is outperformed by the diversified



	@ 5	@ 10	@ 20
Rel	0.621	0.589	0.612
Orig	0.690	0.672	0.698
LM	0.712	0.676	0.695
LDA	0.719	0.685	0.695
Opt	0.746	0.707	0.730

Fig. 6 Comparison of $D_{KL}(U||Q)$ values (left) and α–nDCG values (right) for the manual evaluation

ranking based on language models lm and topic models lda, which in turn are outperformed by the optimal ranking $opti$ based on Wikipedia. This indicates that our proposed evaluation criterion for diversification, which does not require manual annotation, corresponds well with the widely used measure α–nDCG based on manual annotations. Moreover, the fact that the optimal reranking achieves the best α–nDCG confirmes the observation of Santamaría et al. [23] that Wikipedia can be effectively used as a target model for diversification, provided that it covers the most prominent aspects of a query.

7 Conclusions and Future Work

We have presented a reranking approach for balancing the top-k results of Web search engines with respect to diversity by minimizing the variance of their underlying language models and topic models. Our extensive evaluation against Wikipedia has demonstrated that the approach effectively achieves a better coverage of the various topics and aspects pertaining to a query. Our evaluation using the TREC data and supplied evaluation framework confirms these findings and validates the presented Wikipedia-based diversity evaluation as an alternative to costly manual diversity assessment.

We further demonstrated that diversification based on language models achieves a slightly better coverage in terms of Wikipedia language models than diversification based on topic models, but topic models accomplish diversification with a significantly lesser amount of reranking.

We are currently developing an online system to rerank on-the-fly based on Latent Dirichlet Allocation. We want to apply result diversification in the context of summarization of search results as well as of events in blogs and newspaper articles. Moreover, we want to experiment with using cross-entropy and Kullback-Leibler divergence directly for reranking search results such that the top-k results provide a representative overview on the complete result. Finally, we also want to develop approaches to diversification and evaluation, which better focus on the topical content of documents.

Acknowledgements This work is partially supported by the EU Large-scale Integrating Project Living-Knowledge[12] (contract no. 231126). We would like to thank Santamaría et al. for providing us with their manually annotated test corpus [23].

[12] http://livingknowledge-project.eu/

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14, New York, NY, USA, 2009. ACM.
2. W. Bi, X. Yu, Y. Liu, F. Guan, Z. Peng, H. Xu, and X. Cheng. Ictnet at web track 2009 diversity task. In *Text REtrieval Conference*, 2009.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.
4. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM.
5. B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1287–1296, New York, NY, USA, 2009. ACM.
6. H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 429–436, New York, NY, USA, 2006. ACM.
7. C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009*, volume Special Publication 500-278. NIST, 2009.
8. C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, New York, NY, USA, 2008. ACM.
9. W. S. Cooper. The formalism of probability theory in IR: a foundation or an encumbrance? In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 242–247, New York, NY, USA, 1994. Springer-Verlag New York, Inc.
10. A. Das Sarma, S. Gollapudi, and S. Ieong. Bypass rates: reducing query abandonment using negative inferences. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 177–185, New York, NY, USA, 2008. ACM.
11. E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl. Divq: Diversification for keyword search over structured databases. In *SIGIR '10: Proceedings of the 33rd annual international ACM SIGIR conference on Research and development in information retrieval*, July 19–23 2010.
12. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.
13. S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 381–390, New York, NY, USA, 2009. ACM.
14. T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, April 2004.
15. J. He, E. Meij, and M. de Rijke. Result diversification based on query-specific cluster ranking. *Journal of the American Society for Information Science and Technology*, 62(3):550–571, 2011.
16. A. Jain, P. Sarda, and J. R. Haritsa. Providing diversity in k-nearest neighbor query results. In *PAKDD'04: Proceedings of the 8th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 404–413, Berlin, Germany, 2004. Springer.
17. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.
18. H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.
19. F. Radlinksi, P. N. Bennet, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance - workshop report. *ACM SIGIR Forum*, 43(2), December 2009.
20. F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 691–692, New York, NY, USA, 2006. ACM.
21. D. Rafiei, K. Bharat, and A. Shukla. Diversifying web search results. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 781–790, New York, NY, USA, 2010. ACM.
22. S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, pages 294–304, 1977.

23. C. Santamaría, J. Gonzalo, and J. Artiles. Wikipedia as sense inventory to improve diversity in web search results. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1357–1366, Uppsala, Sweden, July 2010. ACL.
24. E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. Amer-Yahia. Efficient computation of diverse query results. In *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, Cancún, México*, pages 228–236. IEEE, 2008.
25. J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122, New York, NY, USA, 2009. ACM.
26. X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, New York, NY, USA, 2006. ACM.
27. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.
28. C. Zhai and J. Lafferty. A risk minimization framework for information retrieval. *Inf. Process. Manage.*, 42(1):31–55, 2006.
29. C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 10–17, New York, NY, USA, 2003. ACM.
30. C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 22–32, New York, NY, USA, 2005. ACM.