# Result Set Diversification in Digital Libraries through the Use of Paper's Claims

José María González Pinto ✉ and Wolf-Tilo Balke

Institut für Informationssysteme
Technische Universität Braunschweig
Braunschweig, Germany
{pinto,balke}@ifis.cs.tu-bs.de

**Abstract.** Understanding the possible associations between two entities from a query is a hard problem. For instance, querying "coffee" and "cancer" even in a curated Digital Library is a challenge to the retrieval system that struggles to figure out the intention of the query. Maybe the user wants a consensus of what it is known? But how many different associations exist? How to find them all? Herein we introduce an approach to diversify the results retrieved from such queries aiming at re-ranking the result list. Our re-ranking models specifically one fundamental aspect of scientific papers: claims. Claims are the sentences that scientists use to report findings. In particular, we study claims that express associations between entities in the medical domain. More specifically, we focus on queries that involve two entities in which one of the entities has some effect on a disease. Thus, we work on a corpus obtained by querying PubMed to empirically assess our proposed solution. Moreover, we promote the idea of claims as an explicit key aspect to consider diversification in the result set of a query. We show the potential of our approach to ease the process of discovering representative associations between entities. Our approach relies on a representation of claims using neural embedding of word vectors and implements an algorithm to perform the re-ranking of the result set of a query. We empirically show the potential of our approach.

**Keywords:** diversification, embedding, scientific claim.

## 1 Introduction

A core functionality of Digital Libraries to satisfy information needs is to provide search capabilities that exploit key aspects of the documents. Delivering high quality results to a query is crucial because of the potential impact of user's decisions. However, as it has been shown in [1] biases are observed during search with respect to two aspects: 1) most of the results support the query while only a few disapprove it and 2) results supporting the query are ranked higher than results contradicting it. Previous work has shown that diversification of the results of a query can alleviate this problem.

However, little attention has been paid to enable such mechanisms to cope with complex information needs in the medical domain where the health conditions of users could be compromised. In particular, when the user is trying to decide about the consumption of a product, a medicine or a drug regarding a specific disease. For this type of information need no doubt that Digital Libraries have better quality content than the Web. For instance, today a user interested in discovering whether a drug is beneficial or not regarding a specific disease, she would have to do an exploratory search submitting several queries. For each query, the user will basically try to get a "consensus" of what the research community has found. Is there a better alternative? In this work, we explore the idea of diversification of the returned set of a given query to help the user in such a task. In particular, we focus on a key aspect of research papers to help the user in her quest: claims. By *claims* in scientific papers we mean statements that express associations between entities. This is of particular relevance in the medical domain where the consumption of a drug, a substance, a fruit, etc., has an effect on a disease. One of the challenges of considering the claims of papers is that the association between two entities can be subject to different *interpretations*. Thus, in this paper, we model a particular case that can arise when interpreting some of the associations between the entities: controversy. One instance of the existence of several controversial claims was found and reported first by [2]. The authors manually discovered, by submitting several queries to PubMed and analyzing the result set relating 50 substances to cancer, that basically most of the substances could increase the risk of cancer and decrease it! The existence of such cases motivates our work to ease the discovery of such cases. Herein, we propose to implement a mechanism to *diversify* the result set of a query to help the user discover entities that may be in a controversial case.

In this work, we aim at modeling the claims of research to perform a re-ranking of the result set of a query represented by two entities. Our approach consists of three basic steps given a pair of <$entity$, $disease$>: firstly, extract from research papers, associations between the pair; secondly, represent the associations using a neural embedding representation of documents and thirdly, deliver a re-ranking of the result set to ease the discovery of controversial claims.

Our proposed approach will bring several benefits: for the information's provider, it will add more value to its current retrieval mechanisms. For the user, the possibility of making an informed decision that can potentially save her life. Moreover, researchers in the medical domain who are in the quest of solving complex problems can also benefit from our approach: they will be able to find controversial claims that basically are in the need of further investigation.

Aiming at this challenge, in this paper, we focus on the design and implementation of a technique that can re-rank documents based on a fundamental aspect of research papers: claims. The remainder of this paper is organized as follows. Section 2 provides definitions and the problem we aim at solving in this paper. Section 3 overviews related work. Sections 4 and 5 describe the experimental setup and the evaluation of our proposed approach. Lastly, Section 6 presents our concluding remarks.

## 2 Model and Problem Definition

In this section, we provide definitions and the problem we aim at solving in this paper. Let's first define what a claim is:

**Definition 1 (Claim):** A claim is a sentence in a research paper that expresses an association between two entities. An association is any verb found in WordNet.

**Definition 2 (Entity):** an entity is the name of a substance, a fruit, vegetable, a drug or a disease.

For example, in the following claim: "lycopene increases the risk of cancer" The entities are "lycopene" and "cancer". The association between the two is "increases".

Now we can define our controversial claim problem below:

**Problem Definition (Claim Diversification Problem)**: Given a collection of $m$ documents $D = \{d_1, \dots, d_m\}$ of a Digital Library, an initial query represented by a pair of entities $< entity, disease >$, we intent to rank documents in $D$ to *diversify* the result set to cover the different interpretations of the associations between$<entity, disease>$ at the top $t$ results.

Our definition resembles the general case [3] where it was proven to be NP-hard in its original form: aiming at maximum coverage with minimum redundancy. However, in our case, we aim at using claims as the proxy to represent an explicit aspect behind the user query instead of the implicit approach that makes the problem NP-hard. Thus, coverage in our work is in terms of the semantics of the associations of claims. And claims are represented as vectors using neural embedding.

We approach the problem by dividing it in the following tasks:

1. Find all the $d \in D$ where a pair of entities $entity, disease$ appear as a claim $< Claims >$. (Section 4.1).
2. Represent each claim $< Claims >$ in an embedding space $< EmbedClaims >$. (Section 4.2).
3. Perform a ranking of the documents using an adaptation of the List of Clusters Diversification algorithm (LCD) originally introduced in [4] and used to accomplish diversification by [5]. (Section 4.3).

In the corresponding sections, we elaborate on the details of each of the tasks. The following section reviews related work.

## 3 Related Work

Our research is related to efforts found in the Web search community towards alleviating biases. Indeed, biases have been a constant problem on the Web and have received considerable attention from different aspects. For instance, in [6] domain bias was investigated in Web search. Domain bias is defined as the user's propensity to believe that a page is more relevant just because it comes from a particular domain. In [1] it was found that users show biases by favoring information that confirms what their beliefs when conducting a search. Researchers proved by a series of experiments the urgent need of search engines to cope with what they called bias and accuracy problem

in the result set of a query. To deal with the problem of bias, several approaches to deliver result diversification have been proposed. These approaches could be categorized as either implicit or explicit [7]. Basically, they differ in how they account for the different query aspects that can help to diversify the result set for a given query. Implicit approaches make the assumption that similar documents will cover similar aspects of the query and should therefore be in the final ranking. The challenge for these methods is to discover the possible different aspects in an unsupervised fashion. A pioneering example presented in [8] introduces a method that basically combines query-relevance with information-novelty in the context of retrieval and summarization. In a similar line of thought in [9] a method was introduced that exploits statistical language modeling to cope with redundancy and relevance. In their work, the problem of sub-topical retrieval is introduced. Basically, the idea is to find documents that cover different subtopics (aspects) of a query. In [10] the use of clustering was introduced to improve the effectiveness of the diversification of the results of a query. Basically, the idea is to first cluster the candidate documents and then restrict the diversified approach to documents associated with clusters that potentially contain many relevant documents. A study comparing implicit diversification techniques with cluster based approaches that select cluster centroids as the representative documents in the final result list is given in [11]. They concluded that clustering is usually a better approach for single subtopics of a given query. However, diversification implicit methods turned out to be better for quick coverage of distinct sub-topics. Another line of research takes diversification with a different perspective. These efforts model specifically the query aspects considered relevant for a specific domain. Usually, some type of external knowledge is exploited to account for these aspects. For instance, in [3] they look at the problem of diversification by assuming that a taxonomy exists. With this assumption, diversification is achieved by favoring documents from different categories and penalizing those that fall into already covered categories. A similar approach is used for product search in [12] where in addition to the categories of products, attributes within each category were considered. In [13] the query aspects were taken from the query log of a commercial search engine. Then, they proposed a ranking to satisfy each aspect of the original query. Another approach that exploits the idea of automatic query reformulations using TREC subtopics is the work of [7]. The researchers introduced a probabilistic approach that explicitly considers the aspects of the query as given by the sub-topics track in the TREC diversification task. The presented approach favors documents that cover those aspects that are not yet covered in the current results set of the generated candidate list. Our work is related to the explicit category of diversification. In our work, we promote claims as first-class citizens and how controversial claims, in particular, can raise in health-related queries.

## 4 Methodology

In this section, we introduce our methods to solve our novel problem of Claim Diversification, to explicitly rank the result set of a query represented as the pair <$entity$, $disease$>.

### 4.1 Dataset

To rely on high quality content, we used PubMed as our main source of documents. For each pair <$entity$, $disease$>, we submitted a query represented as the following query pattern in PubMed:

*(help AND prevent) OR (lower AND risk) OR (increase OR increment AND risk) OR (decrease OR diminish AND risk) OR (factor AND risk) OR (associated AND risk) AND (entity AND disease).*

The ranking provided from PubMed' retrieval system is our initial set of ranked documents $D$. However, not all the documents retrieved from the query were used in our experiments. The main reason was that we wanted to be sure that a claim corresponds to the main contribution of a paper. Thus, we proceeded as follows: firstly, we filtered out documents with no conclusions metadata. Secondly, we split each document in sentences. And thirdly, for each sentence in each document, we selected as the claim of the document the sentence that contained <$entity$, $disease$>. This preprocessing step had a positive impact in the quality of the documents that we used.

### 4.2 Claim Representation

In this section, we provide details of how we represent claims of research papers and how we compute similarity between them for our proposed re-ranking mechanism. To represent the sentence with the <$entity$, $disease$> pair, we used neural embedding of words. Following the success of word embedding representations that capture meaningful semantic relations between words from large text corpus, we opted to represent the claims using word2vec [14, 15]. One particular property that makes this representation useful for our task is that it has been demonstrated that not only are words with similar meanings embedded nearby, but also natural word arithmetic can be applied.

**Claim embedding representation.** Concretely, we represent each claim as the set of the word2vec representation of its words. For our experiments, we relied on the word2vec vectors trained on a combination of all publication abstracts from PubMed and all full-text documents from the PubMed Central Open Access subset [16]. As detailed from the authors, word2vec was run using the skip-gram model with a window size of 5, hierarchical softmax training, and a frequent word subsampling threshold of 0.001 to create 200-dimensional vectors. Another possible representation of words, Glove [17] could also be used for our particular problem.

**Distance metric.** Computing the distance between claims is a fundamental step for our proposed re-ranking mechanism. We decided to use the Word Mover's Distance (WMD) [18] after previous experimentation. As stated by the authors, the WMD distance measures the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to "travel" to reach the embedded words of another document. The proposed WMD was shown to deliver very successful results on document classification data. For our problem, we contrasted it with the cosine similarity and report only our results using WMD because it was superior.

### 4.3 List of Clusters Diversification (LCD)

The idea of List of Clusters Diversification was first introduced in [4]. Basically, the approach relies on the List of Clusters (LC) data structure. LC has been shown to be efficient in high-dimensional metric space searches [5]. In the following paragraphs we include a summary of the explanation of [5]. The idea of the algorithm is to build clusters $(c, r)$. Each cluster has a center $c$ with a covering radius $r$, so that documents in the cluster are within the covering radius of the center.

To diversify a ranking of documents that were initially retrieved from a query, we first need to choose a center $c$ and a radius $r$. The cluster $(c, r)$ comprises the subset of documents of D which are at distance of at most r from c. We define:

$$I_{D,c,r} = \{d \in D\{c\}: \delta(c, d) \leq r\} (1)$$

as the set of internal documents, i.e., which lie inside the cluster $(c, r)$, and

$$\varepsilon_{D,c,r} = \{d \in D\{c\}: \delta(c, d) > r\} (2)$$

as the set of external documents. Clustering is applied recursively in the external set. The function $\delta$ in our case is WMD. The algorithm ends when all documents have been assigned to a cluster. Afterwards, the centers are promoted to the top of the ranking. Furthermore, a center chosen first has preference over the subsequent ones. After that, the remaining of the documents are returned in the order given by its internal membership with respect to its corresponding center. More formally, Algorithm 1 shows how to compute the List of Clusters Diversification (LCD).

---

$\boldsymbol{LCD}[q, D = \{d_1, \dots, d_n\}, k]$

1. $C \leftarrow \{d_1\}$
2. $\varepsilon \leftarrow D\{d_1\}$
3. $c \leftarrow d_1$ #current center
4. while $|\varepsilon| > 0$ do
5. for each $d_i \in \varepsilon$ do
   a.$V = V \cup \{\delta(d_i, c)\}$
6. end for
7. Sort $V$
8. $r \leftarrow V[k]$
9. $I \leftarrow \{d_j \in D\{c\}: \delta(c, d_j) \leq r\} (3)$
10. $\varepsilon \leftarrow \varepsilon \backslash I$
11. $c \leftarrow \{d_i \in \varepsilon\, \delta(c, d_i) > \delta(c, d_j) \forall d_j \in \epsilon\}$
12. $C \leftarrow C \cup \{c\}$
13. $\varepsilon \leftarrow \varepsilon\{c\}$
14. end while
15. $C \leftarrow C \cup \{D \backslash C\}$

---

**Algorithm 1**. List of Clusters Diversification (LCD)

Let's clarify two important aspects of the algorithm. Firstly, the selection of cluster centers. The algorithm uses a ranked list of results and takes as the first cluster the top result. After that, to select cluster centers, line 11 of the algorithm, in [4] was extensively investigated using different heuristics. Experimentally, it was shown that the best strategy is to choose the next center as the object that maximizes the sum of distances to the previous centers. We used in our work the same heuristic. Secondly, the parameter $k$ of the algorithm is used to set the size of the clusters. Empirically, it has been shown that when working with high dimensional metric spaces, the value of $k$ can be dynamically increased as many documents may have the same distance to the center. This is helpful because the number of computations required to select the centers of the clusters can be dramatically reduced. Using a large value of $k$ can help alleviate the cost of distance computations. In our experiments, we set $k$ to six after evaluating a range of values and manually assessing the tradeoffs of the computational cost versus diversity of the result set.

## 5 Experiments

We are aware that the TREC09 and TREC10 collections [19], provide data samples and queries related to the diversification problem in Information Retrieval. Unfortunately, no such data is available for the novel problem presented in this paper where claims are first class citizens. Thus, to evaluate our results, we conducted a series of experiments by querying PubMed as indicated in Section 4.1. Moreover, we propose to use as a metric of our evaluation the Entropy at the top $t$ documents to measure the amount of information expressed in the documents at each $t$. Basically, the idea is that if we achieve higher diversification than the initial result set delivered by PubMed, then we should have a higher entropy. In other words, our proposed method should more evenly divide its probability mass across the documents. Thus, a lower entropy would imply narrow focus of the result set (bias). More formally, entropy is defined as:

$$H(X) = -\sum_{i=1}^{m} p(x_i) \log p(x_i) \quad (4)$$

We performed experiments with 16 entities related to cancer: wine, tea, sugar, salt, potato, pork, onion, olive, milk, lycopene, lemon, egg, coffee, cigar, beef and bacon. We selected these entities for our analysis taken from the cases studied by [2]. We begin in the following paragraphs with a brief discussion of the three main cases found among the 16 entities we analyzed. More specifically, we explain three entities that reflect our main findings: tea, wine and coffee.

In Figures 1, 2 and 3 we plot the entropies at the top 5, 10, 15 and 20 result set with three queries representing three different entities related to cancer: tea, wine and coffee. The label "no diversification" in the plots means the retrieved list of documents where our approach is not used. The label "with diversification" is the one that corresponds to our proposed approach.

The first case, tea and cancer are shown in Figure 1. We can observe that when diversification is applied there is a constant positive difference with respect to the default result set. According to our hypothesis, when diversification is applied up to the top 20 results the user could be better informed.

The second case shown in Figure 2 corresponds to wine and cancer. As it can be observed, it is a different situation: up to the top 10 results our approach could potentially help the user to be aware of a broader set of associations between the entities. However, beginning at the top 15 the differences can be neglected.

In Figure 3 we have the case of coffee and cancer. It seems that our approach is able to diversify the result set. In this particular case, the differences between our approach and the default result set remain constant.

In summary, what we learned from these preliminary experiments is that up to the top 10 results diversification makes a different for this type of data. Even though the differences look small, please notice that our preprocessing step cleaned a lot of data. Because of this preprocessing, the differences do not seem to be as relevant as they could have been expected.

**Comparison with MMR.** To further validate our proposed solution, we also considered in our work the diversity-based re-ranking method called Maximal Margin Relevance (MMR) [8]. We proceeded as follows: we used two metrics to evaluate the differences between the two methods using top 10 results. Firstly, we used entropy as before. And secondly, we computed correlation of word frequencies between each method and the first 10 results with "no diversification". The idea behind this metric is simple but powerful: the performance of one method is worse than the other, the more correlated is with the set of "no diversification". In this work, we used Pearson's correlation with 95 confidence intervals.

To our surprise, the differences between the two methods when using entropy as our main metric are not statistically significant. In Figure 4 we observe the comparisons with each entity and there is no clear winner: in some cases, MMR is better but in half of them LCD does a better job.

However, when we computed the correlation of word frequencies between each method and the top 10 results with no diversification, LCD turned out to be slightly better. In particular, it outperformed MMR in 10 out of the 16 entities.

**Discussion**. One limitation of our current analysis is that qualitatively speaking, we cannot evaluate our approach. We can only observe some differences using entropy as our metric in favor of the idea of allowing a user to get a better overview of a result set. Nevertheless, this is a rather complicated and interesting query type and further work is needed to overcome our current limitations. On the other hand, we could manually observe examples where our approach seems promising. Consider for instance the following top 5 results of our approach for the pair <tea, cancer>:

1. "over consumption of fish sauce, pickled food, moldy cereals, irregularly taking meals and familial history of malignancy may be the local risk factors for high occurrence of gastric cancer, and fresh vegetables and fruits, green tea may have protective effects on it"
2. "our results did not show a protective role of tea in five major cancers"

3. “tea consumption protects against oral cancer in non-smokers or non-alcohol drinkers, but this effect may be obscured in smokers or alcohol drinkers”
4. drinking hot tea, a habit common in golestan province, was strongly associated with a higher risk of esophageal cancer
5. “we observed evidence to support a potential beneficial influence for breast cancer associated with moderate levels of tea consumption (three or more cups per day) among younger women”.

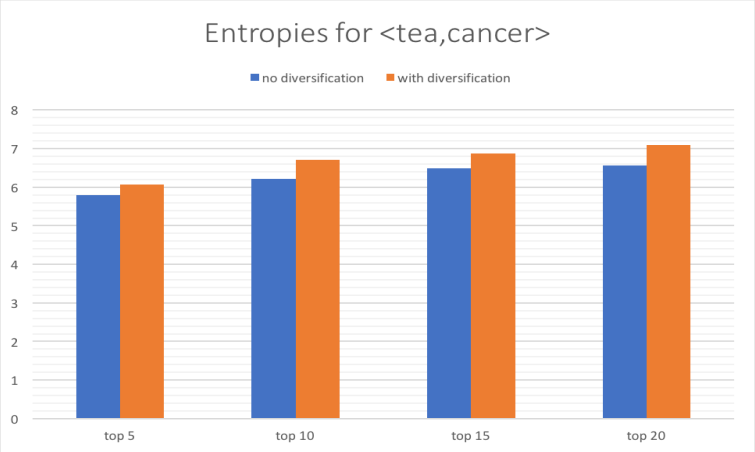


**Fig. 1.** Entropies for the query <tea, cancer>

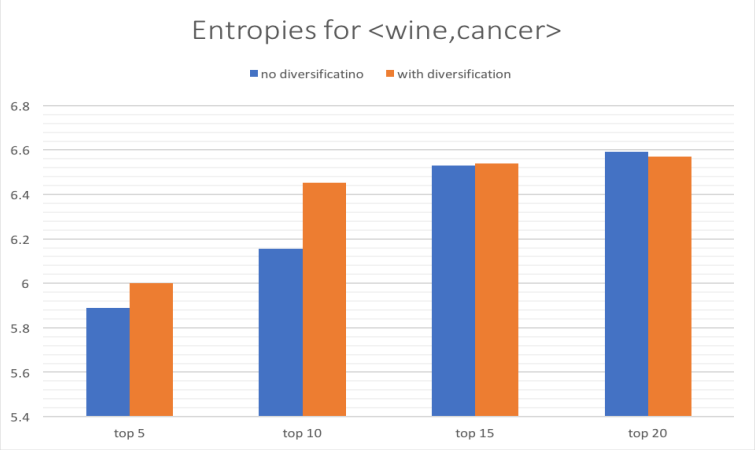


**Fig. 2.** Entropies for the query <wine, cancer>

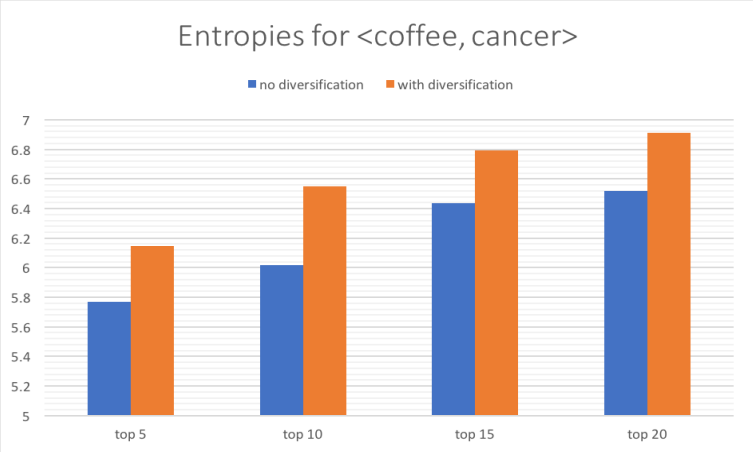


**Fig. 3.** Entropies for the query <coffee, cancer>

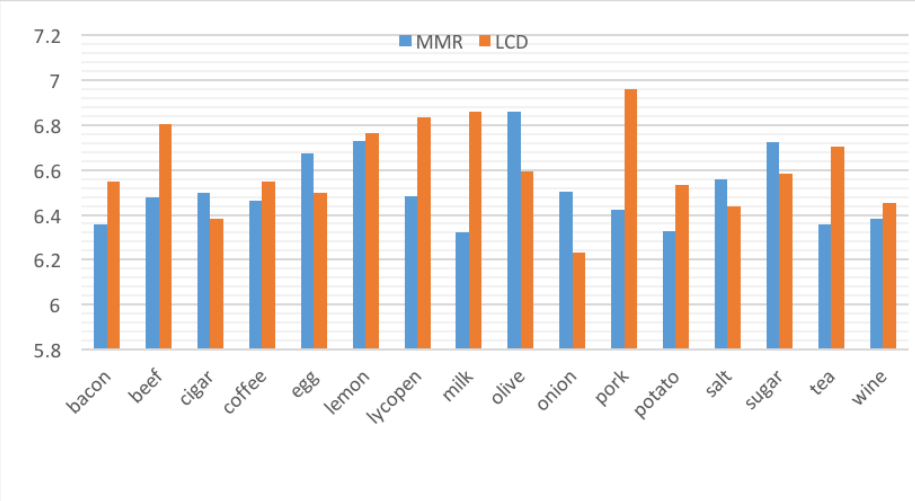


**Fig. 4.** Entropies of MMR (left bar in each pair) and the LCD model (right bar in each pair)

## 6 Conclusions and Future Work

We motivated and presented the novel Claim Diversification Problem for Digital Libraries. In particular, for queries in the medical domain where one entity (a substance, a drug, a medicine, a product, etc.) has some influence with respect to a disease. We build on previous work on Web search where diversification was introduced to deal with the bias on the result set with complex ambiguous queries. In our case, we model specifically one key aspect of scientific papers: claims. Claims in this work are the sentences used in medical research papers to assess the association between two entities.

Our results look promising, and we envision future work to specifically assess the value of promoting claims as the text snippets to present to users from real world queries. Furthermore, we would like to validate the diversification approach that we proposed in this paper with user's feedback. Moreover, we would like to improve our current approach to account for more complex cases where the claims involve more than two entities. Currently, we do not support this type of queries. To accomplish such a task, we would investigate more sophisticated models of the Natural Language community to extract and represent semantically these cases.

We also believe that "time" in the medical domain should be considered as a relevant factor in the diversification process. Therefore, we will incorporate this important factor in our work.

## 7 References


1. White, R.: Beliefs and biases in web search. Proc. 36th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '13. 3 (2013).
2. Schoenfeld, J.D.: Is everything we eat associated with cancer ? A systematic. Am. J. Clinincal Nutr. 97, 127–134 (2013).
3. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09. p. 5 (2009).
4. Chávez, E., Navarro, G.: A compact space decomposition for effective metric indexing. Pattern Recognit. Lett. 26, 1363–1376 (2005).
5. Gil-Costa, V., Santos, R.L.T., MacDonald, C., Ounis, I.: Modelling efficient novelty-based search result diversification in metric spaces. In: Journal of Discrete Algorithms. pp. 75–88 (2013).
6. Ieong, S., Mishra, N., Sadikov, E., Zhang, L.: Domain Bias in Web Search. WSDM '12 Proc. fifth ACM Int. Conf. Web search data Min. 413–422 (2012).
7. Santos, R.L.T.T., Macdonald, C., Ounis, I.: Exploiting Query Reformulations for Web Search Result Diversification. Proc. 19th Int. Conf. World Wide Web. 881–890 (2010).
8. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98. pp. 335–336 (1998).
9. Zhai, C.X., Cohen, W.W., Lafferty, J.: Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. Informaion Retr. 10–17 (2003).
10. He, J., Meij, E., De Rijke, M.: Result diversification based on query-specific cluster ranking. J. Am. Soc. Inf. Sci. Technol. 62, 550–571 (2011).
11. Carpineto, C., D'Amico, M., Romano, G.: Evaluating subtopic retrieval methods: Clustering versus diversification of search results. Inf. Process. Manag. 48, 358–373 (2012).
12. Chen, X., Wang, H., Sun, X., Pan, J., Yu, Y.: Diversifying Product Search

Results. SIGIR. 1093–1094 (2011).
13. Radlinski, F., Dumais, S.: Improving personalized web search using result diversification. Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. SIGIR 06. 691 (2006).
14. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Proc. Int. Conf. Learn. Represent. (ICLR 2013). 1–12 (2013).
15. Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents. Int. Conf. Mach. Learn. - ICML 2014. 32, 1188–1196 (2014).
16. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., Ananiadou, S.: Distributional Semantics Resources for Biomedical Text Processing. In: Proceedings of LBM 2013 (2013).
17. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014).
18. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From Word Embeddings To Document Distances. Proc. 32nd Int. Conf. Mach. Learn. 37, 957–966 (2015).
19. Hawking, D.: Overview of the TREC-9 web track. In: NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9). pp. 87–102 (2001).
20. Manning, C.D., Raghavan, P.: An Introduction to Information Retrieval, http://dspace.cusat.ac.in/dspace/handle/123456789/2538, (2009).