informationresearch

PUBLISHED QUARTERLY BY THE UNIVERSITY OF BORÅS, SWEDEN

VOL. 24 NO. 1, MARCH, 2019

- Contents |
- Author index |
- Subject index |
- Search |
- Home

Evaluating the effectiveness of Web search engines on results diversification

Shengli Wu, Zhongmin Zhang and Chunlin Xu.

Introduction. Recently, the problem of diversification of search results has attracted a lot of attention in the information retrieval and Web search research community. For multi-faceted or ambiguous queries, a search engine is generally favoured if it is able to identify relevant documents on a wider range of different aspects.
Method. We evaluate the performance of three major Web search engines: Google, Bing and Ask manually using 200 multi-faceted or ambiguous queries from TREC.
Analysis. Both classical metrics and intent-aware metrics are used to evaluate search results.
Results. Experimental results show that on average Bing and Google are comparable and Ask is slightly worse than the former two. However, Ask does very well in one subtype of queries – ambiguous queries. The average performance of the three search engines is better than the average of the top two runs submitted to the TREC web diversity task in 2009-2012.
Conclusions. Generally, all three Web search engines do well, this indicates that all of them must use state-of-the-art technology to support the diversification of search results.

Introduction

Since they came into existence on the Web in 1993, Web search engines have been very successful and growing at a very fast speed. Now they are used by billions of people all over the world. During the last two decades, change has been the only constant: new Web search engines have come into existence as some others have faded away.

Shortly after it was launched in 1998, Google took the leading position and held on to it ever since. Its top competitors include Bing, Yahoo!, and Baidu. As of August 2016, Google had a market share of 71.11%, followed by Bing, Baidu, Yahoo!, and Ask with market shares of 10.56%, 8.73%, 7.52%, and 0.24%, respectively (Web search engine).

Of course, every Web search engine company is concerned with both its own performance and that of its competitors. However, these companies do not share their evaluations with the general research community. Recently the problem of search results diversification has attracted a lot of attention in the information retrieval and Web search research community, and many algorithms have been proposed to tackle it. The main objective of the research is to find how leading commercial Web search engines perform in this aspect and gauge the extent to which they are able to satisfy users' information requirements, especially for multi-faceted, ambiguous queries, for which both the relevance and diversity of documents are imperative.

Although a fairly substantial body of research (Lewandowski, 2015; Uyar, 2009; Vakkari, 2011; Wu and Li, 2004) has been given to evaluate the effectiveness of Web search engines employing a range of metrics, to the best of our knowledge, no study has yet broached the topic of results diversification. Thus we anticipate the results from this investigation to be useful to researchers, developers, and users of Web search engines alike. More specifically, we would like to provide an answer to the following two questions:

1. For those multi-faceted, ambiguous queries, how do the leading Web search services perform?
2. Compared with the research community, how does the industry respond to the requirement of search results diversification?

The rest of the paper is organised as follows: first we review existing literature on evaluating the performance of Web search engines, as well as the earlier research in producing diversified retrieval results. Thereafter, we present the investigative methodology and results respectively, while the final section offers the conclusion.

Literature review

Two topics are especially pertinent as the precursors to our study: the evaluation of the effectiveness of Web search engines, and how to enable Web search engines to produce diversified search results.

Search effectiveness evaluation

Early test-based evaluation of Web search engines date back to the 1990s (Chu and Rosenthal, 1996; Ding and Marchionini, 1996; Wishard, 1998; Gordon and Pathak, 1999; Hawking, Craswell, Thistlewaite and Harman, 1999; Leighton and Srivastava, 1999). For example, Leighton and Srivastava (1999) compared the precision of five search engines (Alta Vista, Excite, Hotbot, Infoseek, and Lycos) on the first twenty results returned for fifteen queries. In addition, evaluation took other forms, including survey-based evaluation (Notess, 1995) and log-based evaluation (Jansen, Spink, Bateman and Saracevic, 1998).

In general, to perform a test-based evaluation, one needs a set of queries. Applying each of the queries to a given Web search engine then enables us to obtain a ranked list of documents. Those documents are judged by human assessors to decide whether they are relevant to the information need. Relevance judgment can be binary (a document is either relevant or non-relevant) or graded (more than two categories of relevance). Finally, a variety of metrics can be used to evaluate the effectiveness of the Web search engine over all the queries.

In the following we only reviewed some work on test-based approaches.

Since 2000, more work on evaluation has been conducted, focusing on a range of aspects. Eastman and Jansen (2003) investigated the effect of Boolean operators such as OR, AND, and other types of operators such as must-appear terms and phrases. Can, Nuray and Sevdik (2004) used an automatic method to evaluate a group of Web search engines in which human judgment was not required. Both Wu and Li (2004) and Kumar and Pavithra (2010) compared the effectiveness of a group of Web search engines and Web meta-search engines. Vakkari (2011) compared Google with the digital Ask-a-Librarian service. Uyar and Karapinar (2016) compared the image search engines of Google and Bing. Apart from English search engines, some other languages such as Germen (Griesbaum, 2004), French (Véronis, 2006), Chinese (Long, Lv, Zhao and Liu, 2007), and Arabic (Tawileh, *et al.*, 2010) were also investigated.

Almost all these studies and others besides Lewandowski (2015), Uyar (2009), Deka and Lahkar (2010), Tian, Chun and Geller (2011), Liu (2011), Goutam and Dwivedi (2012) and Balabantaray, Swain, and Sahoo (2013) involve Google, which tends to outperform the other Web search engines under evaluation. One exception is in Vakkari (2011) wherein the investigator finds that Google is outperformed by the Ask-a-Librarian service for queries inferred from factual and topical requests in that service.

Search results diversification

As users grow accustomed to using Web search engines, their needs evolve and increasingly require a diversified set of search results in addition to a relevant set of search results (in particular, those that involve a high degree of duplication). An ambiguous or multi-faceted query is especially difficult for Web search engines to handle, since they may not be able to accurately predict the information need of the user. Personalisation may partially alleviate the problem, but a more general solution is to provide a diversified set of documents, increasing the chances that the user will find relevant and useful information. A recent review on different aspects of result diversification techniques can be found in Abid, *et al.* (2016).

A two-step procedure is usually used to support results diversification when implementing a Web search engine: for a given query, the search engine runs a typical ranking algorithm to obtain a ranked list of documents, considering only relevance; then a result diversification algorithm is applied to re-rank the initial list so as to improve diversity.

A number of approaches have been proposed to diversify search results. We may divide them into two categories: implicit and explicit. When re-ranking the documents, an implicit method does not need any extra information, apart from the documents themselves retrieved through a traditional search system, and possibly some statistics of the whole document collection searched. Carbonell and Goldstein (1998) proposed a maximal, marginal-relevance-based method, which re-ranks documents according to a linear combination of each document's relevance to the query and the similarity between a document and other documents already in the list.

Based on the same idea as Carbonell and Goldstein (1998), Zhai, Cohen and Lafferty (2003) used KL-divergence (Kullback–Leibler divergence) to measure the distance of a new document to those that are already in the list; and both Rafiei, Bharat and Shukla (2010) and Wang and Zhu (2009) used correlation to measure the novelty of a new document to those already in the list. Some methods extract potential subtopics by analysing the documents obtained from the first stage, and then re-ranking them. Analysis can be done in different ways. Carterette and Chandar (2009) extracted potential subtopics by topic modelling, while He, Meij and Rijke (2011) did this by query-specific clustering.

The explicit approach requires more information than the implicit approach. Assuming it is known that the given query has a set of subtopics and other related information, the result diversification algorithm maximizes the coverage of all subtopics in the top-n results. Algorithms that belong to this category include IA-select (Intent-Aware select) (Agrawal, Gollapudi, Halverson and Ieong, 2009), xQuAD (Santos, Macdonald and Ounis, 2010), and proportionality models (Dang and Croft, 2012). The explicit approach is better than the implicit approach if reasonably good information can be collected for different aspects of a given query. Different from the above works which based on a flat list of subtopics, Hu, Dou, Wang, Sakai, and Wen (2015) investigated search result diversification based on hierarchical subtopics.

It is also possible to use a combination of techniques to achieve diversification. In Zheng and Fang (2013), two representative result diversification methods were used, these were xQuAD (eXplicit Query Aspect Diversification) (Santos, Macdonald and Ounis, 2010) and one of the proportionality models – PM2 (Proportionality Model 2) (Dang and Croft, 2012). Liang, Ren and Rijke (2014) presented another combined approach, which comprises classical data fusion, latent subtopics inference, and results diversification. A learning-based approach has also been used for this (Xu, Xia, Lan, Guo and Cheng, 2017; Jiang *et al.*, 2017).

How to evaluate the effectiveness of diversified search results has been investigated by a number of researchers. See Yu, Jatowt, Blanco, Joho and Jose (2017), Wang, Dou, Sakai, and Wen (2016), Chandar and Carterette (2013) for some recent work among many others.

As we can see that in the information retrieval research community, search results diversification has been identified as an important issue and extensive research has been conducted to deal with it. It is interesting to find out how industry responds to this problem and this is the objective of our study in this paper.

Investigation methods

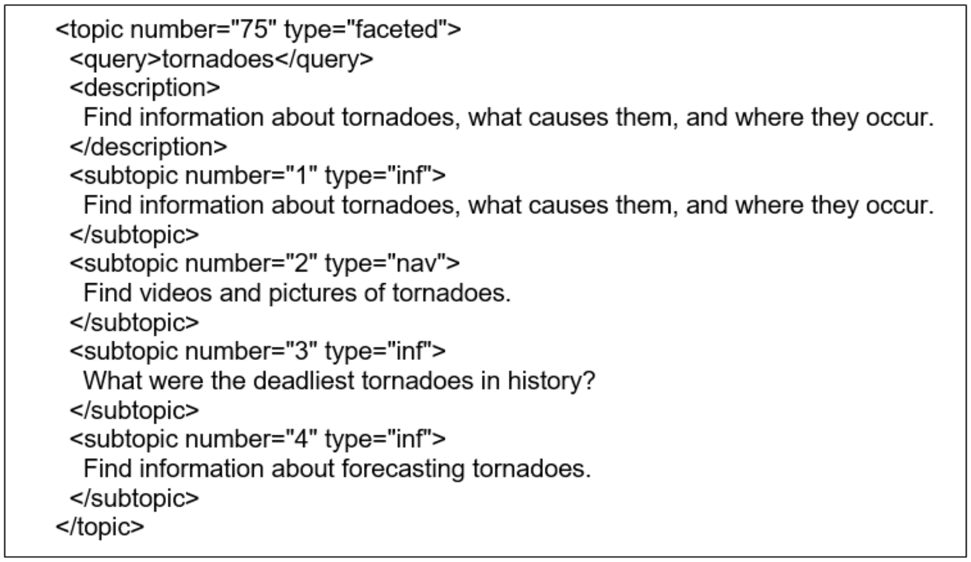
In this study, we investigated three commercial Web search engines Google, Bing, andAsk. Baidu is not included because it does not support English search and Yahoo! is not included because it is powered by Bing.

Eight graduate students undertook the judgment work. To minimize the impact of inconsistencies from different reviewers, each student was allocated an equal number of queries and evaluates all three search engines with all the allocated queries. As another measure for better consistency, we chose ten example queries and let all eight reviewers evaluate the results from all three search engines. The Web documents involved and the judged relevance to the topic were compared and discussed among all reviewers. We hope in this way the same threshold could be leant by all the reviewers for them to carry out their judgment work. In order to avoid the effect of personalized search from Web search engines, each reviewer used a newly-installed Web browser without any search history and used it exclusively for this experiment. For each query, the top twenty documents returned from a search engine were evaluated. The entire process lasted for about four months from the beginning of November 2015 to the end of February 2016. We notice that the search results may vary over time but assume that the performance of a search engine stays the same during the period of testing time.

Query set

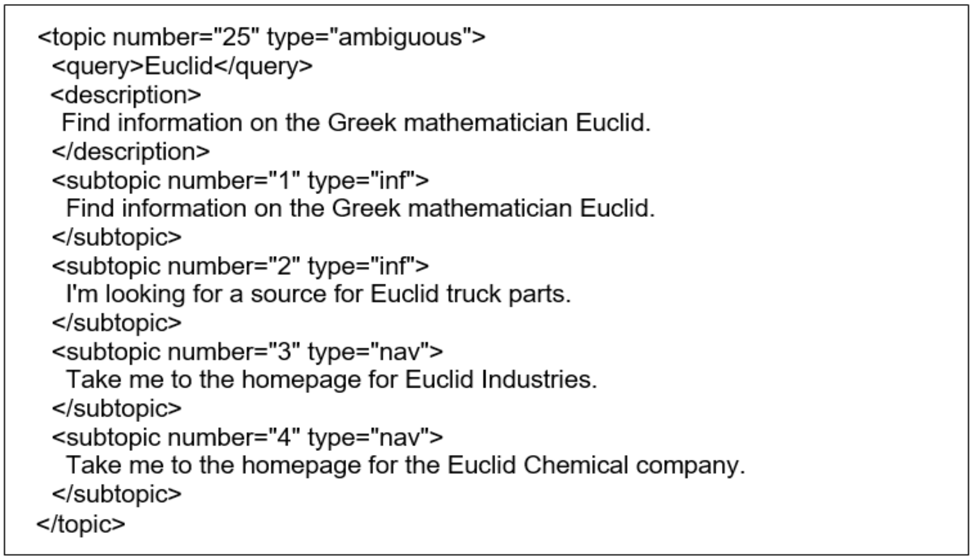
The query set we use is the 200 queries (or topics) that were used in the TREC (Text REtrieval Conference) Web diversity task between 2009 and 2012. The Text REtrieval Conference is a major venue for information retrieval and Web search evaluation. All the queries are categorized into two types: "ambiguous" or "faceted". According to TREC (Clarke, Craswell and Soboroff, 2009), ambiguous queries are those that have multiple distinct interpretations. It is assumed that a user is interested in only one of the interpretations. On the other hand, faceted queries are general and include some related subtopics. A user interested in one subtopic may still be interested in others.

Figure 1 is an example of a faceted query (Query 75) and Figure 2 is an example of an ambiguous query (Query 25) used in TREC.



```
<topic number="75" type="faceted">
  <query>tornadoes</query>
  <description>
    Find information about tornadoes, what causes them, and where they occur.
  </description>
  <subtopic number="1" type="inf">
    Find information about tornadoes, what causes them, and where they occur.
  </subtopic>
  <subtopic number="2" type="nav">
    Find videos and pictures of tornadoes.
  </subtopic>
  <subtopic number="3" type="inf">
    What were the deadliest tornadoes in history?
  </subtopic>
  <subtopic number="4" type="inf">
    Find information about forecasting tornadoes.
  </subtopic>
</topic>
```

Figure 1: A faceted query (Query 75)



```
<topic number="25" type="ambiguous">
  <query>Euclid</query>
  <description>
   Find information on the Greek mathematician Euclid.
  </description>
  <subtopic number="1" type="inf">
    Find information on the Greek mathematician Euclid.
  </subtopic>
  <subtopic number="2" type="inf">
    I'm looking for a source for Euclid truck parts.
  </subtopic>
  <subtopic number="3" type="nav">
    Take me to the homepage for Euclid Industries.
  </subtopic>
  <subtopic number="4" type="nav">
    Take me to the homepage for the Euclid Chemical company.
  </subtopic>
</topic>
```

Figure 2: An ambiguous query (Query 25)

Query 75 includes four subtopics, with three being informational and one being navigational. Query 25 also includes four subtopics, with two being informational subtopics and the two others being navigational. In reality, such queries may have more interpretations, so the subtopics listed are not necessarily complete.

When carrying out the evaluation, the content between tags is used as query input to Web search engines.

Relevance judgment

Binary relevance judgment is used. That is to say, any document is deemed to be either relevant or non-relevant to the information need (a query topic or one of its subtopics). Of course, the same document may be relevant to multiple subtopics of a single query simultaneously. If a document is relevant to one of the subtopics, then this document is regarded as relevant to the query though the converse is not necessarily true: a document can be relevant to a query but without being relevant to any of its subtopics. Several metrics including ERR-IA@m (Intent-Aware Expected Reciprocal Rank at retrieval depth m), P-IA@m (Intent-Aware Precision at retrieval depth m), SubtopicRecall@m, P@m (Precision at document depth m and m = 5, 10, or 20), and MRR (Mean Reciprocal Rank) are used to evaluate search results. ERR-IA@m and P-IA@m are typical metrics for intent-aware evaluation, while P@m and MRR are typical metrics for classical evaluation. P@m is the percentage of relevant documents in all m-top ranked documents. MRR takes the reciprocal of the rank at which the first relevant document appears in the results list.

For a given topic, assume there are n subtopics associated with it. Let R(i, j) = 1 if the document at ranking position *i* is judged relevant to subtopic *j*; otherwise, let R(i, j) = 0. P-IA@m (intent-aware precision) is defined as (Clarke, Craswell, and Soboroff, 2009):

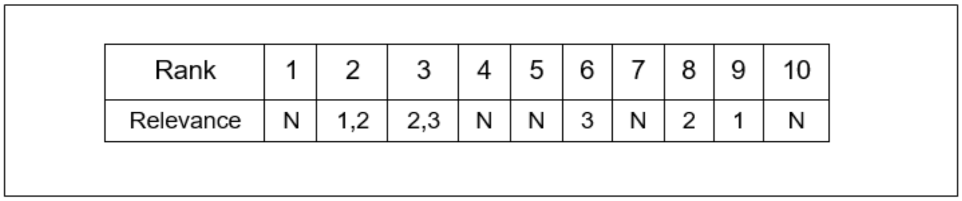
$$P - IA@m = \frac{1}{mn}\sum_{t=1}^{n}\sum_{i=1}^{m} R(i,t) \qquad (1)$$

ERR-IA@m (intent-aware expected reciprocal rank) is defined as (Chapelle, Metzler, Zhang and Grinspan, 2009):

$$ERR - IA@m = \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{m}\frac{1}{k}\prod_{i=1}^{k-1}\left(1-\frac{R(i,t)}{2}\right)\frac{R(k,t)}{2} \qquad (2)$$

Generally speaking, ERR-IA@m is concerned with the total number of documents that are relevant to subtopics up to position m. If a document *d* is relevant to *k* subtopics at the same time, then *d* is counted *k* times. ERR-IA@m favours those results in which all subtopics are covered by the top-ranked documents. ERR-IA@m inspects each of the subtopics separately. According to Clarke, Craswell and Soboroff (2009), SubtopicRecall@m (subtopic recall) can be computed as the number of subtopics with relevant documents in the top *m* divided by the total number of subtopics with relevant documents in the collection. For any subtopic, the ranking position of the first relevant document is the most important factor. Except the first relevant document, all other relevant documents contribute much less to the final score of an intent-aware metric.

Example 1. Suppose for a given query, there are 3 subtopics. The relevance of a results list R to these subtopics is shown in Figure 3.



Rank	1	2	3	4	5	6	7	8	9	10
Relevance	N	1,2	2,3	N	N	3	N	2	1	N

Figure 3: Distribution of relevant documents to all subtopics in an exemplary query

In Figure 3, the number of subtopics to which it is relevant is given, whereas N indicates that the document is not relevant to any subtopics. In the top-5 documents, two documents (at rank 2 and 3) are relevant to at least one subtopic; In the top-10 documents, five documents (at rank 2, 3, 6, 8, and 9) are relevant to at least one subtopic. Therefore, P@5=2/5 and P@10=5/10=1/2. The first document that is relevant to any subtopic is at rank 2, thus MRR=1/2. $P - IA@5 = 1/(3*5)\sum_{i=1}^{3}\sum_{t=1}^{5} R(i,t) = 4/15$, and $P - IA@10 = 1/(3*10)\sum_{i=1}^{3}\sum_{t=1}^{10} R(i,t) = 7/30$. In order to calculate ERR-IA@m, we look at each subtopic separately. For subtopic 1, the relevant documents are at rank 2 and 9. The score that we obtain for this part is 1/3(1/2*1* 1/2+1/9*1/2*1/2). Similarly, the scores for subtopics 2 and 3 are 1/3(1/2*1* 1/2+1/3*1/2*1/2+1/8*1/4*1/2) and 1/3(1/3*1* 1/2+1/6*1/2*1/2), respectively. Summing together these three parts, we produce ERR-IA@5=0.2500 and ERR-IA@10=0.2600.

In all the cases, broken links are treated as non-relevant documents to any topics.

Results

We discuss the evaluation process in two separate parts. In the first part we only consider those subtopics that are listed in TREC. In the second part, we remove this restriction. For any document, if the reviewer thinks it is relevant to a subtopic that is not listed in TREC, then the subtopic will be added to the list of subtopics for that query.

Evaluation using listed subtopics in TREC

The evaluation results are shown in Tables 1 (for Metrics ERR-IA@m and P-IA@m) and 2 (for Metrics P@m and MMR). From Tables 1 and 2, we can see that Bing outperforms both Google and Ask, while Google stands better than Ask in all but one metric: when P-IA@5 is used, Ask is 5.35% more effective than Google. For intent-aware metrics, the difference between them is always small, being below 10% for all cases. For classical metrics, the difference between them is large though the difference is still below 20% for every case. Paired T test is carried out to decide if the difference is significant. In most cases, the difference between Bing and Ask is significant at the level of .05. The difference between Google and Bing is significant on four measures P-IA@5, P@5, P@10, and P@20, but not for seven others. The difference between Google and Ask is significant on six measures.

All the queries can be divided into two categories: ambiguous queries and faceted queries. Search performance is calculated separately for the three search engines (see Tables 3 and 4). We find an interesting phenomenon: Bing is the best performer with faceted queries, while Ask is the best performer with ambiguous queries. Because there are many more faceted queries (142) than ambiguous queries (58), Bing takes with the lead over Ask in our experiment (see Tables 3 and 4). However, should we treat these two types of queries equally without considering the number of queries in each type, then Ask and Bing would be very close on all diversity-based metrics.

In summary, in aggregate Bing performs the best, Google is in second place, while Ask comes last when we consider average performance. Ask performs the best for the ambiguous queries but perform the worst for the faceted queries, while the difference between Bing and Google is small.

Next we look at those metrics more closely. In Table 1, SubtopicRecall@20 is above 0.95 for both Google and Bing, which shows that over 95% of all the subtopics identified in TREC are covered by the top twenty documents. The figure for Ask is 0.8295, which means that about 83% of all the subtopics identified in TREC are covered by the top twenty documents in Ask's results. ERR-IA@m also favours those documents that cover more subtopics. Let us assume there are *n* subtopics for a given query. Then in a results list, the first document relevant to a specific subtopic contributes 0.5/n, and the second relevant to the same subtopic contributes 0.25/n, …, to the final score. Considering the top-*m* documents in a results list, if there is just one relevant document for each of the subtopics, then the final score of ERR-IA@m is 0.5. In Table 1, all ERR-IA@m scores are over 0.4 and two of them are over 0.5. This confirms that most subtopics are covered in the top-*m* documents (for *m*=5, 10, or 20). When more documents are considered, more subtopics are covered for any results list. This is why the value of ERR-IA@m increases with *m*. However, the value of ERR-IA@m increases quite slowly when *m* increases from 5 to 10 and to 20. This shows that a relatively large percentage of subtopics are already covered by five top-ranked documents. On the other hand, P-IA@m only concerns how many documents are relevant to any of the subtopics. Thus the value of P-IA@m decreases when *m* increases.

Another aspect that can be looked at is the difference between ambiguous queries and faceted queries. For each ambiguous query, on average 107.5 documents are relevant to a subtopic (if a document is relevant to *m* subtopics, then it is counted *m* times); while the number of documents is 155.5 for faceted queries. Also for faceted queries, more documents are relevant to more than one subtopic than for ambiguous queries. More specifically, in all 142 faceted queries, 9186, 4391, 1632, 534, 105, and 2 documents are relevant to 1, 2, 3, 4, 5, and 6 subtopics, respectively; in all 58 faceted queries, 5116, 830, 217, 63, and 5 documents are relevant to 1, 2, 3, 4, and 5 subtopics, respectively. This should have an effect on ERR-IA@m and P-IA@m. From Tables 3 and 4, we can see that for both Google and Bing, ERR-IA@m and P-IA@m values are larger for faceted queries than for ambiguous queries. However, Ask is opposite to Google and Bing on this aspect.

Table 1: Intent-aware performance (measured by ERR-IA@m, P-IA@m, and SubtopicRecall@20) of three Web

search engines.

Metric	Google	Bing	Ask	Average
ERR-IA@5	0.4495	**0.4796(6.70%)**	0.4397(-2.18%)+	0.4563
ERR-IA@10	0.4765	**0.5059(6.17%)**	0.4641(-2.60%)+	0.4822
ERR-IA@20	0.4907	**0.5183(5.62%)**	0.4775(-2.69%)+	0.4955
P-IA@5	0.3078	**0.3357(9.06%)***	0.3115(1.20%)+	0.3183
P-IA@10	0.2817	**0.2970(5.43%)**	0.2600(-7.70%)#+	0.2796
P-IA@20	0.2410	**0.2554(5.98%)**	0.2079(-13.73%)#+	0.2348
SubtopicRecall@20	0.9500	**0.9709(2.20%)**	0.8295(-12.65%)#+	0.9167

In all 200 queries, the figures in parentheses are the differences when compared with Google; the figures in bold denote the best performances among three search engines for a given metric; significant difference between two search engines at .05 level is marked by *, #, or +: * for Google vs. Bing, # for Google vs. Ask, and + for Bing vs. Ask

Table 2: Classical performance (measured by P@5, P@10, P@20, and MRR) of three Web search engines.

Metric	Google	Bing	Ask	Average
P@5	0.6880	**0.7740(12.50%)***	0.6440(-6.40%)#+	0.7020
P@10	0.6390	**0.7030(10.01%)***	0.5690(-10.95%)#+	0.6370
P@20	0.5660	**0.6360(12.37%)***	0.4715(-16.69%)#+	0.5578
MRR	0.9156	**0.9558(4.39%)**	0.8803(-3.86%)+	0.9174
Number of broken links	123	**93(-24.39%)**	375(204.87%)	197

In all 200 queries, the figures in parentheses are the differences when compared with Google; the figures in bold denote the best performances among three search engines for a given metric; significant difference between two search engines at .05 level is marked by *, #, or +: * for Google vs. Bing, # for Google vs. Ask, + for Bing vs. Ask

Table 3: Intent-aware performance (measured by ERR-IA@m and P-IA@m) of three Web search engines.

Metric	Google	Bing	Ask
ERR-IA@5	0.4263	0.4340(1.81%)	**0.4570(7.20%)**
ERR-IA@10	0.4495	0.4576(1.80%)	**0.4795(6.67%)**
ERR-IA@20	0.4677	0.4715(0.81%)	**0.4902(4.81%)**
P-IA@5	0.2451	0.2650(8.12%)	**0.3364(37.25%)**
P-IA@10	0.2094	0.2252(7.55%)	**0.2780(32.76%)**
P-IA@20	0.1782	0.1879(5.44%)	**0.2180(22.33%)**

Fifty-eight ambiguous queries; the figures in bold denote the best performances among three search engines for a given metric.

Table 4: Intent-aware performance (measured by ERR-IA@X and P-IA@X) of three Web search engines.

Metric	Google	Bing	Ask
ERR-IA@5	0.4590	**0.4982(8.54%)**	0.4327(-5.73%)
ERR-IA@10	0.4875	**0.5256(7.82%)**	0.4578(-6.09%)
ERR-IA@20	0.5002	**0.5375(7.46%)**	0.4723(-5.58%)
P-IA@5	0.3334	**0.3645(9.33%)**	0.3013(-9.63%)
P-IA@10	0.3113	**0.3264(4.85%)**	0.2527(-18.82%)

P-IA@20	0.2666	**0.2830(6.15%)**	0.2038(-23.56%)

142 faceted queries; the figures in bold denote the best performances among three search engines for a given metric.

Table 5: Performance of the top-two submissions (measured by ERR-IA@20) to the TREC Web diversity task 2009-2012.

Metric	**2009**		**2010**		**2011**		**2012**		**Average**
	Top1	**Top2**	**Top1**	**Top2**	**Top1**	**Top2**	**Top1**	**Top2**	
ERR-IA@5	0.2226	0.1844	0.3103	0.3115	0.4976	0.4651	0.4736	0.4051	0.3588
ERR-IA@10	0.2420	0.2019	0.3343	0.3335	0.5207	0.4897	0.4944	0.4222	0.3798
ERR-IA@20	0.2501	0.2144	0.3473	0.3457	0.5284	0.4994	0.5048	0.4315	0.3902
P-IA@5	0.1619	0.1267	0.2135	0.2108	0.3715	0.3330	0.4080	0.3544	0.2725
P-IA@10	0.1444	0.1124	0.1950	0.1825	0.3521	0.3273	0.3921	0.3414	0.2559
P-IA@20	0.1224	0.1080	0.1703	0.1766	0.3039	0.2910	0.3504	0.3178	0.2301
SubtopicRecall@20	0.7742	0.9248	0.7210	0.8276	0.9756	0.9800	0.9603	0.9508	0.8893
P@5	0.4680	0.3720	0.4042	0.4375	0.4320	0.3600	0.5280	0.4360	0.4297
P@10	0.4040	0.3300	0.3771	0.3771	0.3940	0.3720	0.5200	0.4420	0.4020
P@20	0.3550	0.3230	0.3438	0.3740	0.3510	0.3350	0.4750	0.4050	0.3702
MRR	0.6551	0.6107	0.6581	0.6523	0.6609	0.5725	0.6920	0.5617	0.6329

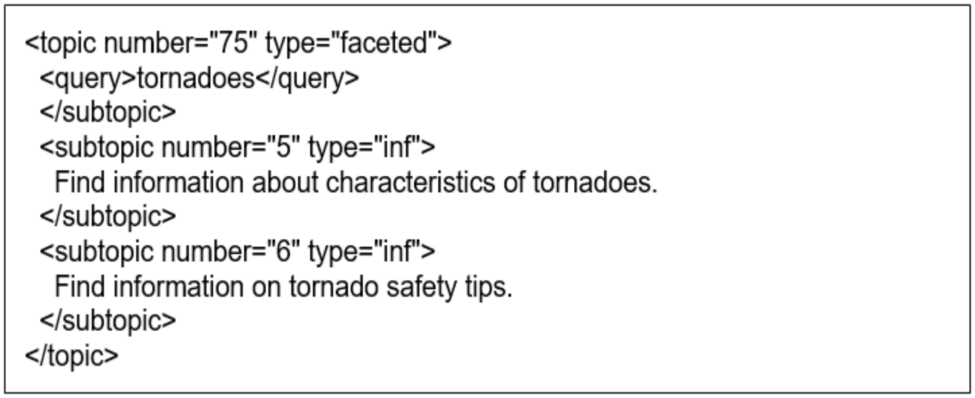
We also select some search results that were submitted to TREC. As the queries were used between 2009 and 2012, we chose the top-two results from each group of those submitted to the TREC Web diversity task in four successive years. The results are shown in Table 5.

From Table 5, we can see that the performance of those runs varies over different years. The two runs in 2009 are the worst, which are followed by the two runs in 2010, and the runs in 2011 and 2012 are close, and all of them are better than those in 2009 and 2010, especially when measured by intent-aware metrics. This phenomenon is understandable because 2009 is the first year in which the Web diversity task was held in TREC, and progress may take place over time. The performance of the submitted runs stabilises in 2012 after improvements in the previous three years. Another factor also contributes to this phenomenon significantly: varying number of sub-topics in those queries. In 2009, the average number of sub-topics is 4.86 per query; while they are 4.36, 3.36, and 3.90 in the next three years. As we will demonstrate later: the more sub-topics a query has, the more difficult it is for retrieval results to achieve better performance if measured by intent-aware metrics. See the next section for detailed analysis and explanation.

The three search engines can be regarded as representatives of the industry and those top-runs submitted to TREC can be regarded as representatives of the academia. By grouping all three commercial search engines and top-two runs submitted to TREC respectively, we compare their average performance over the same group of 200 queries. Comparing the figures in Tables 1 and 2 and the figures in Table 5 (all appear in the last column), we find that all the figures in Tables 1 and 2 are higher than the figures in Table 5. It shows that the industry sector has done a better job than the academia sector. Considering the collection of documents indexed by the Web search engines is considerably greater in both scale and diversity, it is more challenging for Web search engines to achieve comparable performance of those runs submitted to TREC. It indicates that the industry must use state-of-the-art technology to support the diversification of search results.

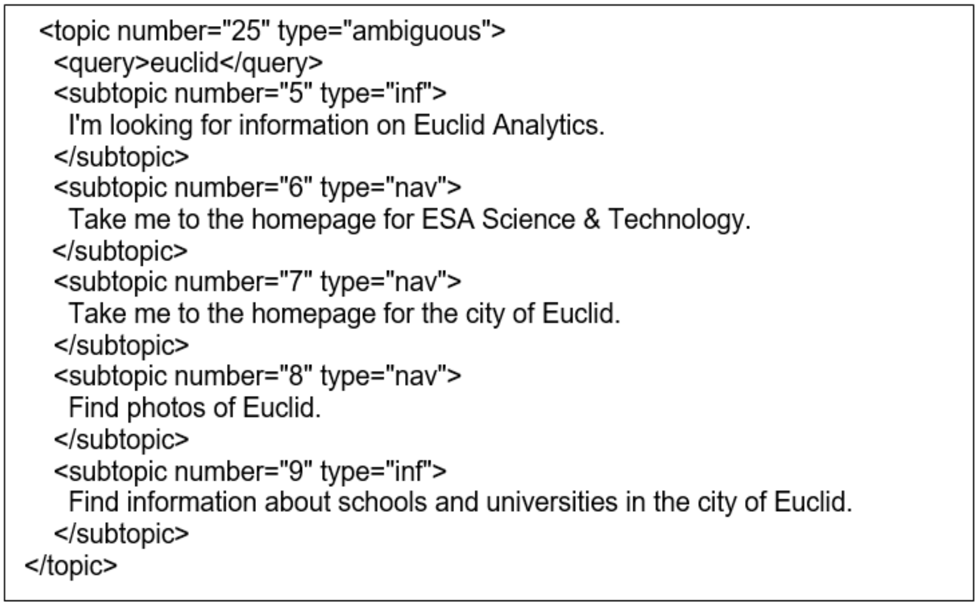
Evaluation using all possible subtopics

In reality, for a given topic, the subtopics provided by TREC may not be complete and some more subtopics exist. Let us consider two queries: Query 75 and Query 25. Query 75 is a faceted query and Query 25 is an ambiguous query. Each of them has four subtopics. When we look at more documents on the Web, two more informational subtopics are found for Query 75 and five more subtopics including two informational and three navigational subtopics are found for Query 25. Figures 4 and 5 show the subtopics added to query 75 and query 25, respectively.



```
<topic number="75" type="faceted">
  <query>tornadoes</query>
  </subtopic>
  <subtopic number="5" type="inf">
    Find information about characteristics of tornadoes.
  </subtopic>
  <subtopic number="6" type="inf">
    Find information on tornado safety tips.
  </subtopic>
</topic>
```

Figure 4: Two more subtopics of Query 75.



```
  <topic number="25" type="ambiguous">
    <query>euclid</query>
    <subtopic number="5" type="inf">
      I'm looking for information on Euclid Analytics.
    </subtopic>
    <subtopic number="6" type="nav">
      Take me to the homepage for ESA Science & Technology.
    </subtopic>
    <subtopic number="7" type="nav">
      Take me to the homepage for the city of Euclid.
    </subtopic>
    <subtopic number="8" type="nav">
      Find photos of Euclid.
    </subtopic>
    <subtopic number="9" type="inf">
      Find information about schools and universities in the city of Euclid.
    </subtopic>
</topic>
```

Figure 5: Four more subtopics of Query 25.

We try to understand what the possible consequence can be if more subtopics are identified. In this part, we look at 20 queries, which are selected randomly from the first 100 queries. They are query 1, 8, 12, 19, 25, 26, 28, 38, 46, 50, 51, 58, 61, 64, 72, 75, 82, 87, 91, and 92. 10 of them are ambiguous queries and 10 are faceted queries. In total, there are 77 subtopics for those 20 queries in TREC and 46 new subtopics have been found in the results from the three Web search engines. This means an increase of 60% over the original 77 subtopics. In other words, each query has an average of 3.85 subtopics initially, which increases by 2.3 subtopics, reaching 6.15. This time we consider all possible subtopics identified both in TREC and in the three commercial Web search engines.

The results are shown in Tables 6 and 7. This time Ask is better than the other two on most diversity-based metrics, while Google is the best on most relevance-based metrics. However, the difference between them is quite small in every case. It seems that the results in Tables 6 and 7 are not very consistent with the results in Tables 1 and 2. This is understandable because this time there are equal number of queries in each category, while in the former case there are far more faceted queries than ambiguous queries.

We also compare the results of the same search engine using subtopics provided by TREC vs. all possible subtopics. We find that, when more subtopics are found in the results list, the values of metrics P@m increase, while the values of intent-aware metrics including ERR-IA@m and P-IA@m decrease. The same phenomenon happens to all three search engines. Figure 6 shows Google's results: paired histograms of different metrics. Decrease rates for ERR-IA@5, ERRIA@10, and ERR-IA@20 are 24.23%, 21.58%, and 20.13%, respectively; the lower rates for P-IA@5, P-IA@10, and P-IA@20 are 22.84%, 17.76%, and 14.07%, respectively; while the higher rates for P@5, P@10, and P@20 are 20.51%, 29.20%, and 34.78%, respectively.

Table 6: Intent-aware performance (measured by ERR-IA@m and P-IA@m) of three Web search engines.

Metric	Google	Bing	Ask
ERR-IA@5	0.3492	0.3702(6.01%)	**0.3961(13.43%)**
ERR-IA@10	0.3774	0.3977(5.38%)	**0.4196(11.18%)**
ERR-IA@20	0.3956	0.4137(4.58%)	**0.4363(10.29%)**
P-IA@5	0.2685	0.2633(-1.94%)	**0.2768(3.09%)**
P-IA@10	0.2405	**0.2461(2.33%)**	0.2399(-0.25%)
P-IA@20	0.2052	**0.2104(2.53%)**	0.2035(-0.83%)

Twenty selected queries with equal numbers of ambiguous and faceted queries, the figures in parentheses are the differences compared to Google; the figures in bold denote the best performance on a given metric

Table 7: Classical performance (measured by P@5, P@10 and P@20) of three Web search engines.

Metric	Google	Bing	Ask
P@5	**0.9337**	0.8600(-7.89%)	0.8500(-8.96%)
P@10	**0.8729**	0.8250(-5.49%)	0.7450(-14.65%)
P@20	0.7593	**0.7600(0.09%)**	0.6625(-12.75%)
MRR	**1.0000**	0.9750(-2.50%)	0.9292(-7.08%)

Twenty selected queries with equal numbers of ambiguous and faceted queries, the figures in parentheses are the differences compared to Google; figures in bold denote the best performance on a given metric.

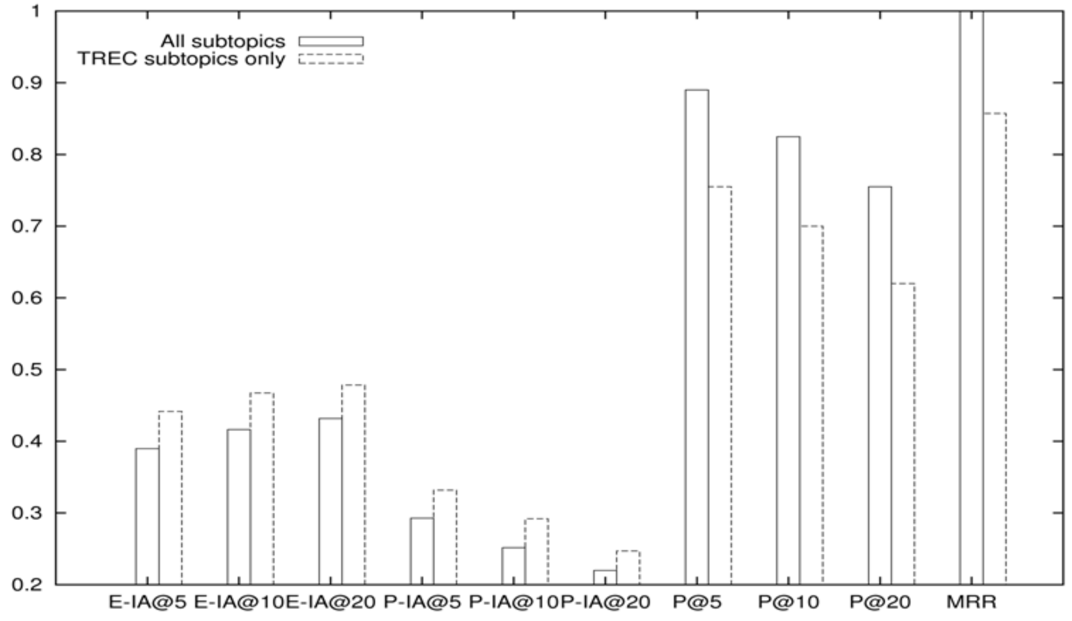


Figure 6: Google's performance: considering TREC-specified subtopics vs. considering all possible subtopics in the resulting list.

To confirm this finding, we further look at all 200 queries but with TREC-listed subtopics only. In all these queries, the minimum number of subtopics specified for a query is 3 while the maximum number is 8. We divide them into two groups: queries with 3-5 subtopics (Group 1) and queries with 6-8 subtopics (Group 2), and separately compare the performance of each of the three search engines in these two groups. Figure 7 shows Bing's results. The same phenomenon also happens to the two other Web search engines.

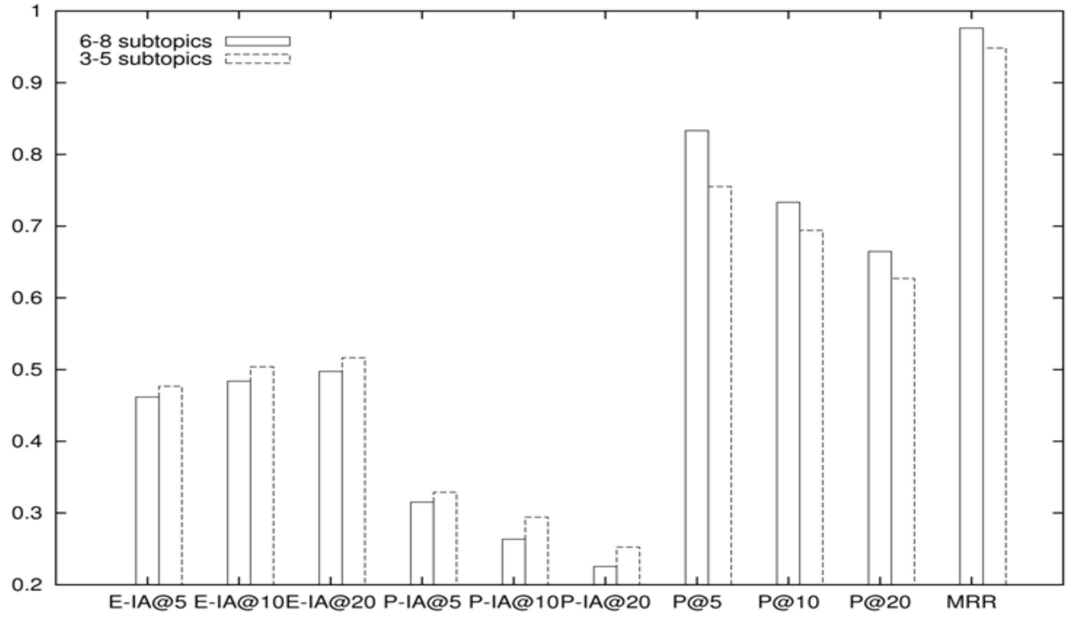


Figure 7: Bing's performance: 6-8 subtopics vs. 3-5 subtopics.

The formulae (Equations 1 and 2) for metrics ERR-IA@m and P-IA@m help to explain the underlying cause of this phenomenon. For any given topic, all subtopics share one resulting list. If we assume that each document is relevant to at most one subtopic, then the more subtopics there are, the more difficult it is for any subtopic to get a relevant document amongst the top ranking positions. Thus the final average of all subtopics will decrease. On the other hand, because a document relevant to any subtopic will be counted as being relevant to the general query, with more identified subtopics the probability that a given document is relevant to at least one of the subtopics increases, meaning that P@m and MMR values will increase accordingly.

This finding is interesting and significant for the research and experiments of search results diversification. If there are two groups of queries and the queries in group A have fewer subtopics than the queries in group B, then it is easier for a search engine to achieve better performance with the queries in B than with the queries in A. Further research work is desirable in two directions of retrieval evaluation with certain overlaps. One is about retrieval evaluation metrics. The intent-aware metrics can be modified to make them consistent over different number of sub-topics. The other is about retrieval evaluation experiments. When designing such experiments, the number of sub-topics should be considered as a factor that has significant impact on the difficulty level of the experiment. Furthermore, if we compare two results from different experiments, then their difficulty levels, including the numbers of sub-topics for the queries involved, should be considered.

Conclusions

In this investigation we have evaluated the performance of three major Web search engines: Google, Bing, and Ask, primarily focusing on their ability to diversify results. Through extensive experimentation, we find that all of them perform well. When considering top twenty documents in the results, then, on average, over 80% of the subtopics are covered by Ask; and over 90% of the subtopics are covered by both Google and Bing. We have also compared the results of these three search engines with the top two results submitted to the TREC Web diversity task between 2009 and 2012 and find that the average performance of the former group is better than the average of the latter group. This indicates that all the search engines support results diversification effectively powered by the state-of-the-art technology.

More specifically, we find that Bing and Google are comparable and both of them are slightly better than Ask on intent-aware metrics. Such a phenomenon is somewhat surprising given that most previous investigations have found that Google is more effective than the others. However, previous investigations have not taken into account results diversification. Furthermore, the queries tested in this investigation may not be the most common queries submitted to the search engines. Thus the results from this investigation reflect one aspect of those search engines, though are unlikely to present the whole picture of user satisfaction, meaning that the results obtained in this investigation do not necessarily conflict with those from previous investigations.

Another finding in this investigation is that the number of subtopics has opposite impact on intent-aware metrics. That is to say, intent-aware metrics favour queries with fewer subtopics. This would recommend further work to make the intent-aware metrics fair to queries with varying numbers of subtopics and performance values comparable over different search result diversification experiments.

About the authors

Shengli Wu is a Professor at the School of Computer Science and Telecommunication Engineering, Jiangsu University, China. His research areas include information retrieval and machine learning. He can be contacted at swu@ujs.edu.cn.
Zhongmin Zhang is a Graduate Student at the School of Computer Science and Telecommunication Engineering, Jiangsu University, China. She can be contacted at shinezzm@foxmail.com.
Chunlin Xu is a Ph D Student in the Computing Department, Ulster University, UK. She can be contacted at xu-c@ujs.edu.cn.

References

- Abid, A., Hussain, N., Abid, K., Ahmad, F., Farooq, M. S., Farooq, U., … & Sabir, N. (2016). A survey on search results diversification techniques. *Neural Computing and Applications 27*(5), 1207-1229.
- Agrawal, R., Gollapudi, S., Halverson, A., & Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (pp. 5-14). New York: ACM.
- Balabantaray, R. C., Swain, M., & Sahoo, B. (2013). Evaluation of web search engines based on ranking of results and features. *International Journal of Human Computer Interaction, 4*(3), 117-127.
-
- Can, F., Nuray, R., & Sevdik, A. B. (2004). Automatic performance evaluation of Web search engines. *Information processing & management, 40*(3), 495-514.
-
- Carbonell, J., & Goldstein, J. (1998). 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* (pp. 335-336). New York: ACM.
- Carterette, B., & Chandar, P. (2009). Probabilistic models of ranking novel documents for faceted topic retrieval. In *Proceedings of the 18th ACM Conference on Information and knowledge management* (pp. 1287-1296). New York: ACM.
- Chandar, P., & Carterette, B. (2013). Preference based evaluation measures for novelty and diversity. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 413-422). New York: ACM.

- Chapelle, O., Metzler, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 621-630). New York: ACM.
- Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: A comparative study and evaluation methodology. *Proceedings of the Annual Meeting-American Society for Information Science, 33*, 127-135.
- Clarke, C. L., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 web track. In *Proceedings of the Eighteenth Text REtrieval Conference.* Natioanal Institute of Standards and Technology, USA. Retrieved from https://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf. (Archived by WebCite® at http://www.webcitation.org/766kAUa8O)
- Dang, V., & Croft, W. B. (2012). Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 65-74). New York: ACM.
- Deka, S. K., & Lahkar, N. (2010). Performance evaluation and comparison of the five most used search engines in retrieving web resources. *Online Information Review, 34*(5), 757-771.
- Ding, W., & Marchionini, G. J. (1996). A comparative study of web search service performance. *Proceedings of the ASIST Annual Meeting, 33*, 136-140.
- Eastman, C. M., & Jansen, B. J. (2003). Coverage, relevance, and ranking: the impact of query operators on Web search engine results. *ACM Transactions on Information Systems, 21*(4), 383-411.
- Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management, 35*(2), 141-180.
- Goutam, R. K., & Dwivedi, S. K. (2012). Performance evaluation of search engines via user efforts measures. *International Journal of Computer Science Issues, 9*(4), 437–442.
- Griesbaum, J. (2004). Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. *Information Research, 9*(4) paper 189. Retrieved from http://www.informationr.net/ir/9-4/paper189.html (Archived by WebCite® at http://www.webcitation.org/7661JHsvu)
- Hawking, D., Craswell, N., Thistlewaite, P., & Harman, D. (1999). Results and challenges in web search evaluation. *Computer Networks, 31*(11), 1321-1330.
- He, J., Meij, E., & de Rijke, M. (2011). Result diversification based on query-specific cluster ranking. *Journal of the American Society for Information Science and Technology, 62*(3), 550-571.
- Hu, S., Dou, Z., Wang, X., Sakai, T., & Wen, J. (2015). Search result diversification based on hierarchical intents. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management* (pp. 63-72). New York: ACM.
- Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: a study of user queries on the web. *ACM SIGIR Forum, 32*(1):5-17. New York: ACM.
- Jiang, Z., Wen, J., Dou, Z., Zhao, W., Nie, J., Yue, M. (2017). Learning to diversify search results via subtopic attention. In *Proceedings of the 40th international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 545-554). New York: ACM.
- Kumar, B. T., & Pavithra, S. M. (2010). Evaluating the searching capabilities of search engines and metasearch engines: a comparative study. *Annals of Library and Information Studies, 57*(2), 87–97.
- Leighton, H. V., & Srivastava, J. (1999). First 20 precision among World Wide Web search services (search engines). *Journal of the Association for Information Science and Technology, 50*(10), 870.
- Lewandowski, D. (2015). Evaluating the retrieval effectiveness of Web search engines using a representative query sample. *Journal of the Association for Information Science and Technology, 66*(9), 1763-1775.
- Liang, S., Ren, Z., & De Rijke, M. (2014). Fusion helps diversification. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 303-312). New York: ACM.
- Liu, B. (2011). User personal evaluation of search engines-Google, Bing and Blekko. *University of Illinois at Chicago*. Retrieved from https://www.cs.uic.edu/~liub/searchEval/Search-Engine-Evaluation-2011.pdf. (Archived by WebCite® at http://www.webcitation.org/76AxxQrxB)
- Long, H., Lv, B., Zhao, T., & Liu, Y. (2007). Evaluate and compare Chinese Internet search engines based on users' experience. In *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing* (pp. 6134-6137). Piscataway, NJ: IEEE.

- Notess, G. R. (1995). Searching the World-Wide Web: Lycos, WebCrawler and more. *Online, 19*(4), 48-53.
- Rafiei, D., Bharat, K., & Shukla, A. (2010). Diversifying web search results. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 781-790). New York: ACM.
- Santos, R. L., Macdonald, C., & Ounis, I. (2010). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 881-890). New York: ACM.
- Tawileh, W., Mandl, T., Griesbaum, J., Atzmueller, M., Benz, D., Hotho, A., & Stumme, G. (2010). Evaluation of five web search engines in Arabic language. In *Proceedings of LWA Workshop* (pp. 221-228). Retrieved from http://www.kde.cs.uni-kassel.de/conf/lwa10/papers/ir1.pdf. (Archived by WebCite® at http://www.webcitation.org/6zfYEBTJa)
- Tian, T., Chun, S. A., & Geller, J. (2011). A prediction model for web search hit counts using word frequencies. *Journal of Information Science, 37*(5), 462-475.
- Uyar, A. (2009). Investigation of the accuracy of search engine hit counts. *Journal of Information Science, 35*(4), 469-480.
- Uyar, A., & Karapinar, R. (2016). Investigating the precision of web image search engines for popular and less popular entities. *Journal of Information Science, 43*(3), 378-392.
- Vakkari, P. (2011). Comparing Google to a digital reference service for answering factual and topical requests by keyword and question queries. *Online Information Review, 35*(6), 928-941.
- Véronis, J. (2006). A comparative study of six search engines. *University of Provence*. Retrieved from https://www.researchgate.net/publication/265028347_A_comparative_study_of_six_search_engines. (Archived by WebCite® at http://www.webcitation.org/6zfhAifXV)
- Wang, J., & Zhu, J. (2009). Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 115-122). New York: ACM.
- Wang, X., Dou, Z., Sakai, T., & Wen, J. (2016). Evaluating search result diversity using intent hierarchies. In *Proceedings of the 39th international ACM SIGIR conference on Research and development in information retrieval* (pp. 415-424). New York: ACM.
- Web search engine. (2016). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Web_search_engine
- Wishard, L. (1998). Precision among Internet search engines: an earth sciences case study. *Issues in science and technology librarianship, 18*(1). Retrieved from http://webdoc.gwdg.de/edoc/aw/ucsb/istl/98-spring/article5.html. (Archived by WebCite® at http://www.webcitation.org/7660sOsRr)
- Wu, S. & Li, J. (2004). Effectiveness evaluation and comparison of Web search engines and meta-search engines. In *International Conference on Web-Age Information Management* (pp. 303-314). Berlin: Springer
- Xu, J., Xia, L., Lan, Y., Guo., J., & Cheng, X. (2017). Directly optimize diversity evaluation measures: a new approach to search result diversification. *ACM Transactions on Intelligent Systems and Technology, 8*(3), Article 43.
- Yu, H., Jatowt, A., Blanco, R., Joho, H., & Jose, J M. (2017). An in-depth study on diversity evaluation: the importance of intrinsic diversity. *Information Processing & Management, 53*(4), 799-813
- Zhai, C. X., Cohen, W. W., & Lafferty, J. (2003). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 10-17). New York: ACM.
- Zheng, W., & Fang, H. (2013). A diagnostic study of search result diversification methods. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval* (p. 17). New York: ACM.

How to cite this paper

Wu, S., Zhang, Z. & Xu, C. (2019). Evaluating the effectiveness of Web search engines on results diversification. *Information Research, 24*(1), paper 811. Retrieved from http://InformationR.net/ir/24-1/paper811.html (Archived by WebCite® at http://www.webcitation.org/76tOkn3pA)

Find other papers on this subject

Scholar Search Google Search Bing

Check for citations, using Google Scholar

Facebook Twitter LinkedIn More

- Contents |
- Author index |
- Subject index |
- Search |
- Home