What Snippets Say About Pages in Federated Web Search

Demeester, T.; Nguyen, D.; Trieschnigg, D.; Develder, C.; Hiemstra, D.

published in
Proceedings of AIRS 2012
2012

document version
Peer reviewed version

What Snippets Say About Pages in Federated Web Search

Thomas Demeester[1], Dong Nguyen[2],
Dolf Trieschnigg[2], Chris Develder[1], and Djoerd Hiemstra[2]

[1] Ghent University, Ghent, Belgium
{tdmeeste,cdvelder}@intec.ugent.be
[2] University of Twente, Enschede, The Netherlands
{d.nguyen,d.trieschnigg,d.hiemstra}@utwente.nl

Abstract. What is the likelihood that a Web page is considered relevant to a query, given the relevance assessment of the corresponding snippet? Using a new federated IR test collection that contains search results from over a hundred search engines on the internet, we are able to investigate such research questions from a global perspective. Our test collection covers the main Web search engines like Google, Yahoo!, and Bing, as well as a number of smaller search engines dedicated to multimedia, shopping, etc., and as such reflects a realistic Web environment. Using a large set of relevance assessments, we are able to investigate the connection between snippet quality and page relevance. The dataset is strongly inhomogeneous, and although the assessors' consistency is shown to be satisfying, care is required when comparing resources. To this end, a number of probabilistic quantities, based on snippet and page relevance, are introduced and evaluated.

Keywords: Web search, test collection, relevance judgments, federated information retrieval, evaluation, snippet.

1 Introduction

Finding our way around among the vast quantities of data on the Web would be unthinkable without the use of Web search engines. Apart from a limited number of very large search engines that constantly crawl the Web for publicly available data, a large amount of smaller and more focused search engines exist, specialized in specific information goals or data types (e.g., online shopping, news, multimedia, social media).

The goal of Federated Information Retrieval (FIR) [1] is to combine multiple existing search engines into a single search system. With the wide variety of existing resources, including those that are not directly accessible by Web crawlers, federated search on the Web has an enormous potential, but is a huge research challenge all the same. A number of FIR research collections have been created in the past, but they are mostly artificial and do not represent the heterogeneous Web environment, i.e., search engines with different retrieval methods,

highly skewed sizes, many types of content, and various ways of composing result snippets.

We created a large dataset for this setting, introduced in Nguyen et al. [2], containing sampled results from 108 search engines on the internet, and relevance judgements for both snippets and pages on a number of test topics. We are convinced that such a dataset can stimulate research on federated Web search, and have therefore made our dataset available to researchers[3].

The goal of this paper is threefold. First, we discuss the relevance judgments for the new dataset. Second, we point out some potential difficulties in Web FIR evaluation due to the non-homogeneous character of the resources. Third, we provide a probabilistic analysis of the relationship between the indicative snippet relevance and the relevance of pages.

After a brief overview of related work, several aspects of the relevance judgments are described, starting with the setup of the test collection and the user study, followed by an analysis of the relevance judgments, and focusing on the relationship between snippets and pages. Finally, some conclusions are formulated.

2 Related Work

FIR has been under investigation for many years. The early work is described in detail by Callan [3], identifying three main tasks. *Resource description* is the task of gathering knowledge about the different collections, *resource selection* deals with selecting a subset of collections that are most likely to return relevant results, and *results merging* deals with integrating the ranked lists from different resources into a single ranked list. An extensive overview of more recent work, including FIR in the Web context, is given by Shokouhi and Li in [1].

Reference data in the form of manually annotated relevance judgments are essential in IR evaluation. These are often being created by means of dedicated assessor panels, e.g., for the Text Retrieval Conference (TREC) collections [4]. An important issue for a reliable evaluation is the consistency of the relevance judgments. Voorhees [5] studied the consistency among different assessors, whereas Scholar [6] looked at the self-consistency of assessors for standard TREC collections. Carterette [7] argued that the evaluation accuracy may be affected by assessor errors, which in turn are more likely to occur when judgments are gathered by means of crowdsourcing. Because of the unknown properties of our collection, we only used reliable assessors.

Most online search engines present their results as a ranked list of snippets, giving a sneak preview of the document's content. Nowadays, these snippets are mostly *query-biased*, i.e., they are extracted from the result document, based on the query terms. The fast generation of result snippets has been studied in the past, e.g. by Turpin [8]. For this paper, we will however look at the snippets from a user's perspective, regardless of the different methods used to create them. To

[3] http://www.snipdex.org/datasets

our knowledge, no recent work has been published that investigates the relation between results snippets from various origins, and page relevance.

3 Relevance Assessments

3.1 Test Collection Setup

Our primary goal was to create a test collection with similar properties as the actual Web. Therefore, our data was crawled from the Web, using results from actual search engines, each with their own document collection and retrieval algorithms. Between December 2011 and January 2012, we collected data from 108 resources, ranging from large general Web search engines to search engines over small, specific collections. These resources cover a broad range of application domains and data formats, and we divided them into 12 categories. Besides General Web Search engines, we have Multimedia, Academic, News, Shopping, Encyclopedia, Jobs, Blogs, Books... and several other types of resources. Due to space limitations, we have to refer to Table 4 for a complete listing, and to Nguyen et al. [2] for some example resources per category.

The collection consists of a large amount of web data, obtained by query-based sampling and intended for resource selection experiments, and also a large set of relevance judgments for a number of test topics, to be used for the evaluation of retrieval algorithms. This paper focuses on the relevance judgment analysis, but more detailed properties of the complete dataset are given in [2].

3.2 User Study Setup

We decided to obtain relevance judgments for the top 10 results that each search engine returned in response to a number of test topics. With over a hundred search engines, more than a thousand pages would need to be judged for a single query. Alternative strategies requiring less judgments have been studied, e.g., by Carterette [9], but we wanted a reliable overview of the relevance for this highly inhomogeneous collection. Hence, we decided for independent graded relevance judgments. Also, the setting of many different real-life search engines provided an excellent opportunity to study, besides the pages, the snippets as generated by these resources. When assessors were presented with snippet and page simultaneously, a preliminary experiment showed that the page influenced their snippet judgment. Hence, a snippet and the corresponding page had to be judged separately and in that order, but still by the same assessor to guarantee an unbiased analysis. We decided to first gather all the snippet judgments, and then the page judgments, as it allowed to minimize the page annotation effort, as explained below.[4]

[4] This however does not allow us to investigate how the knowledge of a snippet influences the corresponding page judgement. In fact, based on the insights gained in this study, we will adapt the design of a future user study in several ways, to be able to study such issues.

In the following, we give an overview of the different aspects of the relevance judgments.

Topics The judged topics are those from the 2010 TREC Web Track [10], as we preferred to use existing topics designed for the Web context, to ensure an objective characterization of our collection. These topics are divided into two categories (*ambiguous* and *faceted*), and provided with one general information need, as well as several specific descriptions (for the Web Track diversity task). We only presented the assessors with the query terms and the general information need. Most of the topics are especially suited for general Web search engines, and therefore we pay extra attention to these in our analysis. However, the goal of the test collection is to include the other resource categories as well. Therefore, the judges were instructed to interprete the information need in a broad ('multimedia') sense. Consider, e.g., the query *joints* (a search for information about joints in the human body). A picture or video fragment of human joints, even without further textual data, could be highly relevant. Of course, it is often not possible to interprete a topic from the point of view of all types of search engines. For example, a job offer for an orthopedic surgeon, although related, could not be considered relevant to the query *joints*.

Snippets For each topic, the snippets were shown one by one and in a random order to the assessors. The title, snippet text, preview (if present), and page url, were displayed for each snippet in a uniform manner, albeit provided by the different web search engines. The goal of the snippet annotation task was to predict whether the corresponding page would be relevant. The following labels were used: No, Unlikely, Maybe, and Sure. The corresponding guidelines are summarized in Table 1. We will call these the *snippet relevance* levels, although strictly speaking they only represent the judge's estimation of the page's relevance given the snippet. Especially the smaller resources often provided less than 10 results per query, and in total only 35.651 snippets had to be judged. About 71% of the snippets were judged only once, but for the snippets of 14 topics we obtained between 2 and 5 judgments, with in total over 53000 snippet judgments being collected.

Pages For the page judgments, the Web Track 2010 relevance levels and descriptions were used: Non, Rel, HRel, Key, Nav, see Table 1. We presented the judges with a snapshot of each page, as well as the html content. Judging pages appeared much more time-consuming than judging snippets, but we were able to reduce the page annotation effort by two thirds, based on the following assumption. In case none of the annotators had labeled a particular snippet higher than No, the corresponding page was not judged, and by default given the label Non. As such, we are not able to estimate the probability of page relevance for totally non-relevant snippets. Yet, even if one of those snippets would correspond to a relevant page, the user would not have clicked the snippet, and therefore not have visited the page.

Table 1: Relevance levels and descriptions for snippets and pages

Snippet judgment guidelines: this snippet's page ...	
No	... is definitely not relevant; you would not click the link
Unlikely	... is probably not relevant; you would not click the link
Maybe	... is probably relevant; you would click the link
Sure	... is definitely relevant; you would click the link
Page judgment guidelines: this page ...	
Non	... provides no useful information on the topic
Rel	... provides minimal information on the topic
HRel	... provides substantial information on the topic
Key	... is dedicated to the topic and is worthy of being a top result
Nav	... represents the intended home page of the query entity

Assessors From the 10 assessors that contributed to the relevance judgments, 3 were external students, who created 55% of the snippet judgments, and 57% of the page judgments. One of them was especially hired for one month, and created almost as many judgments as all the other judges together. Apart from the external students, we had 4 junior IR researchers judge 39% of snippets and pages, and 3 senior IR researchers, who did 6% of the snippets and 4% of the pages. We decided to limit the annotation effort for the pages to 39 topics to be judged once and 11 topics twice. The judgments that will be used throughout this paper, are those for which one and the same person has judged all snippets and all required pages of a full topic, i.e., those 39 topics with a single judge, and 11 topics with two judges. Except for the experiments where 2 assessors are compared, we will include those 11 topics in our analysis by randomly selecting the judgments of only one assessor.

3.3 Relevance Judgment Consistency

In the following paragraphs, we will take a closer look at the consistency of the relevance judgments. For clarity, we introduce the random variables S and P, indicating the snippet label, respectively, page label, taking the values listed in Table 1. In some experiments, we reduced our graded relevance levels to a binary relevance, using different cut-offs. For example, the relevance cut-off S$\geq$Maybe means all snippets with labels Maybe and Sure are considered relevant, and the cut-off P$\geq$Key indicates that only pages with labels Key and Nav are considered relevant.

Consistency of Page Judgments per Assessor Some of the resources (especially general Web search engines) often returned the same urls. This allows to study how consistently each assessor performed the page judgments. In [6], assessor consistency on standard TREC collections was investigated by comparing the relevance for pairs of duplicate documents where at least one document had been judged relevant. There were fractions of 15% to 19% of inconsistent

judgments for binary relevance, and between 19% and 24% for ternary relevance. We repeated this experiment, for all (4644) pairs of duplicate urls judged by the same user and with at least one judgment level of Rel or higher. For our five graded relevance levels, we found 24% of differently judged pairs, and for binary relevance judgments (i.e., relevant for labels HRel and above) we found 13%, comparable with the TREC judgments.

In order to obtain consistent page judgments for our analysis further on, we grouped all the judgments for a particular url by the same user, and replaced them with the average judgment level.

Inter-Assessor Consistency The consistency among pairs of assessors has been investigated for a standard TREC collection [5]. It appeared that despite a considerable disagreement among assessors, the ranking of different IR systems mostly remains independent of which set of relevance judgments is used.

We compared our assessors by splitting up our double assessments into two sets (called set 1 and set 2), using binary relevance to allow comparing with [5]. The results are shown in Table 2. The first shown parameter is the *overlap*, calculated as the size of the intersection of the relevant documents in set 1 and set 2, divided by the size of its union, and averaged over the 11 considered topics. In the case of binary page relevance for labels P≥HRel, the resulting overlap of 0.43 (taking into account all resources) is similar to [5], where mean overlap values from 0.42 to 0.49 were reported. From the overlap at different relevance levels, it appears to be more difficult for an assessor to choose between two higher labels (e.g., HRel and Key), than between two lower labels (such as Non and Rel).

We also calculated the fraction of relevant documents from set 1 that are also labeled as relevant in set 2, and vice versa. The average of these two values is shown in Table 2, for convenience called the *precision*[5]. The result at relevance level P≥HRel, including all resources, is 0.62. This imposes a practical upper bound of 62% precision at a recall of 62% on the performance of retrieval systems that are evaluated with these data, because the assessors only agree up to that level. In [5], a value of 65% was reported over the TREC-4 topics. However, our results are strongly influenced by the different types of resources. For the general Web search engines (WSE) on their own, the consistency between judges is much higher than without WSE, as shown in Table 2. This might be due to the fact that the information needs become less well-defined when interpreted in a broader sense than only for general Web search. Also, for the weaker relevance criterion P≥Rel, comparable to the relevance criterion for old TREC collections, we find a much higher precision of 78%.

In order to get an idea of the consistency in judging snippets, we included the snippet overlap and precision in Table 2. Comparing the snippet level S=Sure with the page level P≥HRel, it appears that for the non-WSE, judging snippets

[5] These values can indeed be interpreted as the average of the precision when considering the relevant entries in set 1 as the retrieved set with set 2 as the reference set and vice versa, or, equivalently the average of the recall, by exchanging set 1 and set 2.

Table 2: Average overlap and precision between assessors over pairwise judged queries, for different binary relevance levels, shown for (1) all resources, (2) only general Web search engines (WSE), and (3) without WSE.

		all resources		only WSE		all but WSE	
	relevance	overlap	precision	overlap	precision	overlap	precision
pages	≥Rel	0.64	0.78	0.81	0.90	0.56	0.71
	≥HRel	**0.43**	**0.62**	**0.59**	**0.74**	**0.31**	**0.51**
	≥Key	0.34	0.46	0.36	0.48	0.25	0.33
snippets	≥Maybe	0.53	0.72	0.70	0.83	0.45	0.65
	Sure	**0.37**	**0.59**	**0.60**	**0.75**	**0.22**	**0.42**

is much more difficult than judging pages. This already became clear during the annotation phase, because of the properties of the content (e.g., videos were reduced to small static previews or not shown at all in the snippets), and because the amount of data the decision is based on, is much smaller in the case of snippets. These are difficulties that make evaluation of FIR in a realistic Web context a difficult task.

3.4 Distribution and Uniqueness of Relevant Results

Relevance distribution Figure 1 gives an overview of the average number of resources per topic that returned a given number of relevant pages, split up into ambiguous and faceted topics. The Nav judgments were merged with the Key judgments, because only a few topics were apt for navigational relevance. On average 8 resources returned at least one Key+Nav result for the ambiguous topics, and 9 resources for the faceted topics. However, only for half of the ambiguous topics there was at most one resource that returned 4 or more Key+Nav results, whereas three resources returned 4 or more Key+Nav results for each of the faceted topics. With respect to the total amount of resources (108), the number of resources that returned relevant results is very small for both types of topics; the relevance distribution is highly skewed, which was to be expected, due to the large variation in size and nature of the resources.

An important characteristic of the test collection is also the distribution of the relevant results among the different types of search engines, see Figure 2. The types of search engines that are best suited for the Web Track topics, are in the first place the WSE, but also search engines that provide information from encyclopedia, books, and blogs, and multimedia search engines.

Uniqueness of Results Table 3 gives the number of snippets and pages with a high relevance level, returned by the WSE and the multimedia resources (as these provide by far the most relevant results), and in total. The 10 WSE together provide more highly relevant results than the 21 multimedia search engines, and even more than the total of 98 non-WSE. However, after normalizing the urls to a uniform form (by omitting search-engine-specific additions etc.) and

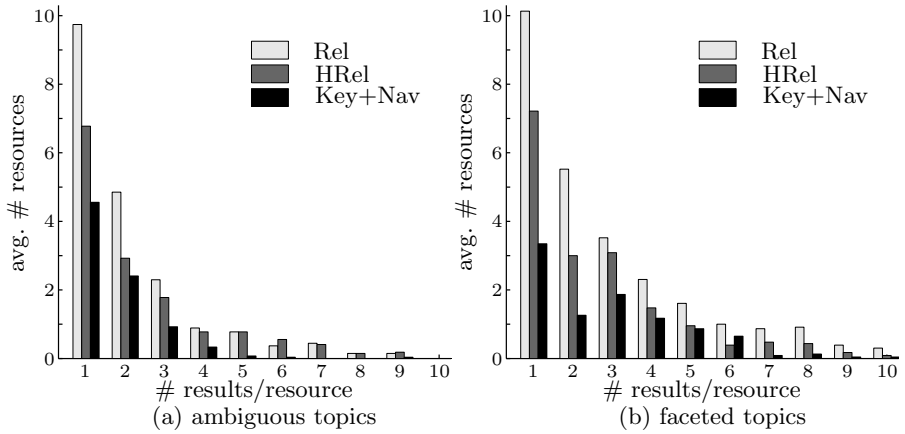

Fig. 1: Average number of resources per topic with a given number of relevant results for the shown relevance levels, for different types of topics.

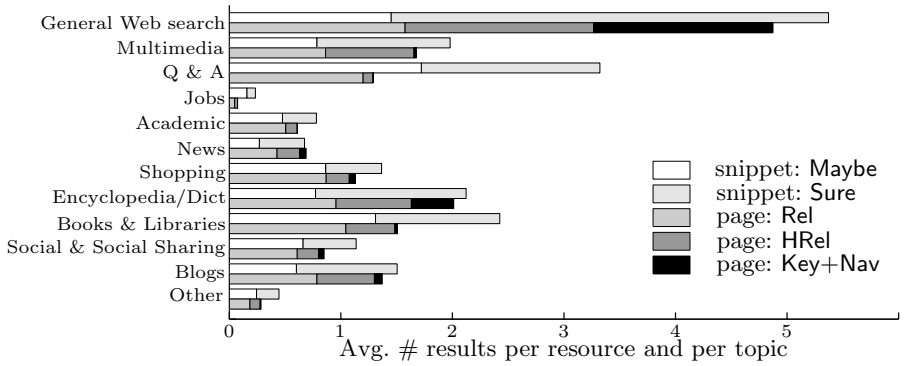

Fig. 2: Average number of resources per topic with a given number of relevant results for the shown relevance levels, for different types of topics.

Table 3: Total number of highly relevant snippets and pages

Resources	# snipp. Sure	# pag. ≥HRel	# unique pag. ≥HRel
General Web Search	1932	1626	675
Multimedia	1038	707	701
Total	4237	2973	1950

counting duplicates only once, the contribution of WSE and multimedia engines is comparable, and in total the non-WSE produced about twice the number of unique highly relevant pages as the WSE. This is a strong argument in favour of FIR on a Web scale. However, when considering the highest relevance level, P≥Key, the WSE engines provide about twice as many unique relevant results compared to the multimedia engines.

3.5 Snippet vs. Page Judgments

This section describes the relation between the relevance labels of snippets and their corresponding pages, judged independently but by the same assessor(s).

Conditional and Joint Probabilities Based on the relevance judgments for all topics, we can make an empirical estimate of the average probability of page relevance given the snippet label, $\mathcal{P}(\mathrm{P}|\mathrm{S})$. The results for different snippet labels and types of resources are presented in Table 4, for binary page relevance P≥HRel.

For snippets with label No, we cannot estimate $\mathcal{P}(\mathrm{P}|\mathrm{S})$, since we have not judged the corresponding pages. However, for most resources, the probability of a HRel page for an Unlikely snippet is already very low. Only for the WSE, 1 out of 5 Unlikely snippets points to a HRel page, suggesting we might have missed a significant number of relevant pages behind non-relevant snippets.

Comparing $\mathcal{P}(\mathrm{P}|\mathrm{S})$ for the snippet labels Maybe and Sure shows that for the most suited types of resources, a relatively large amount of HRel pages are behind snippets which were judged only Maybe. This shows that often a HRel page's snippet cannot convince the user that the page is indeed highly relevant.

For snippets labeled Sure, we showed $\mathcal{P}(\mathrm{P}|\mathrm{S})$, $\mathcal{P}(\mathrm{P},\mathrm{S})$, and $\mathcal{P}(\mathrm{P})$. The page relevance probability $\mathcal{P}(\mathrm{P})$ by itself gives a false impression, as it does not encorporate any effects of the snippet quality. Instead, we could use $\mathcal{P}(\mathrm{P},\mathrm{S})$, denoting for S=Sure and P≥HRel the joint probability that a snippet is judged as Sure *and* its page is at least HRel, but for most of the listed resource categories it is very low, just like $\mathcal{P}(\mathrm{P})$. We can conclude no more than that these collections probably contain little relevant information on our test topics. This is why we instead consider $\mathcal{P}(\mathrm{P}|\mathrm{S})$. It only allows studying how accurately the snippets reflect the page relevance, but is less dependent on the scope of the test topics. For example, the News resources display a relatively high $\mathcal{P}(\mathrm{P}|\mathrm{S})$, against a very low $\mathcal{P}(\mathrm{P},\mathrm{S})$. In other words, they returned only very few relevant results for our topics, but if a snippet is found relevant, 4 out of 10 times it points to one of those few relevant results (see Table 4).

Assessor Dependency An important question is how the relationship $\mathcal{P}(\mathrm{P}|\mathrm{S})$ between snippets and pages would generalize among different users. For example, in a federated search scenario where snippet lists are cached locally and shared among peers, what is the probability that, if a snippet is considered relevant by one user, the corresponding page is relevant to other users? Using the 11 topics with double judgments, we calculated $\mathcal{P}_{\mathrm{cross}}(\mathrm{P}|\mathrm{S})$, the estimated probability of page label P from one assessor given snippet label S from the other assessor (averaged over both directions). In Table 5, these values are shown for different levels of page relevance, together with the average self-probability $\mathcal{P}_{self}$ per assessor. The estimated probability that the page of a relevant snippet is also relevant, would be expected to decrease, if both are not judged by the same assessor. Alternatively, the probability that a page is still relevant, despite a less

Table 4: Overview of the relationship between page and snippet judgments, for different types of resources, and based on the page relevance level P≥HRel.

	S=Unlikely	S=Maybe	S=Sure		
	$\mathcal{P}$(P|S)	$\mathcal{P}$(P|S)	$\mathcal{P}$(P|S)	$\mathcal{P}$(P,S)	$\mathcal{P}(P)$
General Web search	0.20	0.40	0.65	0.26	0.34
Multimedia	0.09	0.23	0.48	0.06	0.09
Q & A	0.00	0.00	0.06	0.01	0.01
Jobs	0.00	0.06	0.24	0.00	0.00
Academic	0.03	0.08	0.14	0.01	0.01
News	0.09	0.19	0.42	0.02	0.03
Shopping	0.06	0.10	0.21	0.01	0.03
Encyclopedia/Dict	0.05	0.23	0.58	0.11	0.14
Books	0.12	0.10	0.18	0.02	0.05
Social & Social Sharing	0.06	0.12	0.19	0.01	0.03
Blogs	0.12	0.23	0.40	0.05	0.07
Other	0.04	0.08	0.34	0.01	0.01
All	0.09	0.21	0.50	0.06	0.08

Table 5: Average $\mathcal{P}$(P|S), for page and snippet judged by different assessors (cross), vs. the same (self).

		S≤Maybe		S=Sure	
		$\mathcal{P}_{\text{self}}$(P|S)	$\mathcal{P}_{\text{cross}}$(P|S)	$\mathcal{P}_{\text{self}}$(P|S)	$\mathcal{P}_{\text{cross}}$(P|S)
General Web Search	P≥HRel	0.16	0.17	0.67	0.65
	P≥Key	0.04	0.08	0.37	0.29
Multimedia	P≥HRel	0.05	0.05	0.49	0.45
	P≥Key	0.00	0.00	0.07	0.04

relevant snippet, is likely to increase in that case. This phenomenon is indeed observed, especially for the stricter page relevance level (P≥Key). This confirms that our precaution, to have one and the same annotator judge corresponding snippets and pages, was necessary.

Comparison between Largest Web Search Engines As the test topics are best suited for the general Web search engines, we can explicitly compare the performance of four of the largest general Web search engines in our collection, i.e., Google, Yahoo!, Bing, and Baidu, as well as Mamma.com, which is actually a metasearch engine. Table 6 presents the results.

For the snippet label S=Sure and two page relevance levels (P≥HRel and P≥Key), we show $\mathcal{P}$(S) and $\mathcal{P}$(P|S), the estimate that a user makes about the page's relevance based on the snippet alone, and how well the page relevance is linked with that estimate. A search engine should however try to optimize the joint probability $\mathcal{P}$(P,S). For a better comparison between these resources, we therefore explicitly report $\mathcal{P}$(P,S) as well. For these resources, we can use $\mathcal{P}$(P,S) as a measure of comparison, because they have a similar target area (i.e., general

Table 6: Comparison of the largest general Web search engines

		P≥HRel and S=Sure			P≥Key and S=Sure		
	$\mathcal{P}$(S=Sure)	$\mathcal{P}$(P|S)	$\mathcal{P}$(P,S)	$\mathcal{P}$(P)	$\mathcal{P}$(P|S)	$\mathcal{P}$(P,S)	$\mathcal{P}$(P)
Google	0.42	0.68	0.28	0.38	0.39	0.16	0.19
Yahoo!	**0.47**	0.69	**0.32**	**0.44**	0.38	0.18	**0.22**
Bing	0.41	0.60	0.24	0.28	0.30	0.12	0.13
Baidu	0.21	0.43	0.09	0.12	0.23	0.05	0.06
Mamma.com	0.43	**0.73**	0.31	0.41	**0.44**	**0.19**	**0.22**

Web search). We also showed the more traditional $\mathcal{P}$(P), which is actually the averaged precision@10 of page relevance, and that is consistently higher than $\mathcal{P}$(P,S), as it does not take the snippet into account.

The metasearch engine outperforms the others, as it aggregates results from a number of resources, such as Google, Yahoo!, and Bing. We want to stress that the considered test topics are still no representative collection of, for example, popular Web queries, and therefore we cannot draw any further conclusions about these search engines beyond the scope of our test collection. Yet, here is another example of how the table might be interpreted, with that in mind. Considering only Key results, we could compare Yahoo! and Bing. Yahoo! seems to score higher for all reported parameters, so either Bing's collection contains a smaller number of relevant results, or Yahoo!'s retrieval algorithms are better tuned for our topics. The lower value of $\mathcal{P}$(P|S) for Bing shows that it has a slightly increased chance that the page for a promising snippet appears less relevant. However, the ratio of $\mathcal{P}$(P,S) and $\mathcal{P}$(P) is higher for Bing than for Yahoo!, indicating that for Yahoo!, its own recall on Key pages will be decreased more due to the quality of the snippets, than for Bing. In fact, we found that $\mathcal{P}$(S=Sure|P≥Key) is 79% for Yahoo!, but 91% for Bing.[6]

4 Conclusions and Future Work

In this paper, we analyzed the relevance judgments for a new federated IR test collection, containing data from over a hundred online search engines with various purposes, such as general Web search, multimedia, academic, books, shopping, etc. It appeared that on average the judgments were created with a level of consistency comparable to that of standard TREC collections, with however a higher consistency for the general Web search engines, as compared to the others. We found the judged test topics from the 2010 Web Track to be biased towards general Web search, leading to a large variation in the fraction of relevant results per resource category. Yet even then, the total number of different relevant results from the non general Web search resources was large enough to confirm the potential interest of federated Web search.

[6] We did not elsewhere elaborate on the parameter $\mathcal{P}$(S|P), due to length constraints. However, the reader could approximate it (due to the low number of digits shown), dividing $\mathcal{P}$(P,S) by $\mathcal{P}$(P) in Tables 4 and 6.

Due to the different scope of the resource categories, an absolute comparison in terms of the empirical probability of relevance for our specific test topics would yield little information. Instead, we discussed the conditional probability of page relevance given the snippet judgment, which allowed a limited but less biased means of comparison. Within the more homogeneous category of general Web search engines, we were able to use the joint probability of snippet and page relevance to compare between the resources. The effect of the snippets varied among the resources, but the joint probability appeared consistently lower than the probability of relevance for only the pages, hence showing that the effect of the snippets cannot be left out when characterizing search engines.

In the future, new query sets will be used, designed to make more use of all the resources under study, including the more specialized ones. The user study will be redesigned, to allow investigating issues like how bad snippets damage recall. We will also investigate the effect of the type and number of topics on the quality of evaluation data for a federated Web search test collection, as well as the potential of resource selection and results merging algorithms.

Acknowledgments. This research was partly supported by the Netherlands Organization for Scientific Research, NWO, grants 639.022.809 and 640.005.002, and partly by the IBBT (Interdisciplinary Institute for Broadband Technology) in Flanders.

References

1. Shokouhi, M., Li, L.: Federated Search. Foundations and Trends in Information Retrieval **5**(1) (2011) 1–102
2. Nguyen, D., Demeester, T., Trieschnigg, D., Hiemstra, D.: Federated Search in the Wild: the Combined Power of over a Hundred Search Engines. Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM) (2012)
3. Callan, J.: Distributed information retrieval. Advances in Information Retrieval **7** (2002) 127–150
4. Voorhees, E.: TREC: Experiment and Evaluation in Information Retrieval. The MIT Press, Cambridge, MA, USA (2005)
5. Voorhees, E.: Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing and Management **36** (2000) 697–716
6. Scholer, F., Turpin, A., Sanderson, M.: Quantifying test collection quality based on the consistency of relevance judgements. In: SIGIR 2011, New York, NY, USA, ACM Press (July 2011) 1063–1072
7. Carterette, B., Soboroff, I.: The effect of assessor error on IR system evaluation. In: SIGIR 2010, ACM (2010) 539–546
8. Turpin, A., Tsegay, Y., Hawking, D., Williams, H.E.: Fast generation of result snippets in web search. In: SIGIR 2007, ACM (2007) 127–134
9. Carterette, B., Allan, J., Sitaraman, R.: Minimal test collections for retrieval evaluation. In: SIGIR 2006, ACM (2006) 268–275
10. Clarke, C.L.A., Craswell, N., Soboroff, I., Cormack, G.V.: Overview of the TREC 2010 Web Track. TREC (2010) 1–9