2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Empirical analysis of the rank distribution of relevant documents in web search

Shen Jiang, Sandra Zilles, and Robert Holte
Department of Computing Science
University of Alberta, Edmonton, Alberta, Canada
{sjiang1,zilles,holte}@cs.ualberta.ca

Abstract

This paper proposes an empirical approach for analyzing the rank distribution of relevant documents in web search. From a methodological point of view a new transaction log analysis method is proposed; the relevance of documents is studied over transaction sessions rather than single transactions. From a practical point of view the paper provides insights about the actual rank distribution of relevant documents in web search, with several consequences for the design of applications related to web search.

1. Motivation and main contributions

Re-ranking algorithms, query refinement and query suggestion methods, document clustering approaches—these and many other techniques are deployed to provide users of a web search engine better access to the documents relevant for their queries in the context of their information need. Many of these techniques assume that whether or not a document is relevant for a query is determined by its rank in the result list for this query. Naturally, one would expect a document to be the more relevant in the context of a query the higher it is ranked in the list of retrieved documents for the same query. This corresponds to a monotonically decreasing distribution of the probability of a document being relevant given its rank. It is reasonable to assume such a distribution; there is, however, *(i)* no evidence concerning the slope of such a distribution and in particular *(ii)* no evidence on the rank at which the probability of being relevant becomes negligible and *(iii)* no evidence as to how big a portion of relevant documents is expected to be found up to this rank.

The goal of this paper is to empirically provide such evidence, more specifically, to determine a probability distribution of relevant documents given their rank.

Such a distribution would first of all hopefully provide a justification for commonly used methods intrinsically assuming certain relations between rank and relevance. Furthermore, such a distribution has applications in the design of search methods. For instance, it could be combined with existing studies on how the rank of a document in a result list determines whether or not a user looks at the document snippet (*i.e.*, how deep the user looks into the list of retrieved documents). A possible application would be to determine, for a given query, a set of documents that are likely to be relevant and that the user is unlikely to have seen. Such a set, sorted by probability of being relevant, could be the target for, *e.g.*, re-ranking or query refinement/suggestion.

The main contribution of this paper is twofold. *Methodologically* it proposes a new transaction log analysis method; *practically* it provides insights about the actual rank distribution of relevant documents in web search.

1.1. Methodological contribution

Our proposed method for determining the desired probability distribution is based on the analysis of search engine transaction logs, aiming at an automated process that will scale to large transaction logs on very large document collections (unlike the rather costly and restricted pooling methods used for evaluation in the TREC series, *cf.* [1]). The problem of determining the ranks of relevant documents cannot be solved by simply adopting methods for collecting statistics over the ranks of documents clicked by the user after issuing a single query—even under the simplifying assumption (a) that all these documents are relevant. The main drawback would not be this simplifying assumption but the fact that its counterpart (b) that *all relevant documents are clicked by the user* is not true. In fact there is evidence that users look only at documents that are very highly ranked, thus missing lower ranked relevant documents, *cf., e.g.,* [11]. Such a statistic would thus be biased.

The approach proposed below partly overcomes this drawback by generalizing both the method and the underlying assumptions. The idea is to collect statistics from user "sessions" rather than from single queries, a session being a sequence of transaction records that form a single user's sequence of interaction with a search engine in the context

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.200

of a single information need. Such sequences can contain several queries. For instance, assume a user issues the following four queries consecutively.

garden plants
ivy garden
edmonton hockey
edmonton oilers nhl

Intuitively, though issued by the same user, these queries should be considered to form two sessions separated between the second and the third query.

The proposed method for determining the desired probability distribution is to

- consider the set D of clicked documents (especially the document d viewed at the end of a session) as relevant for the initial query q_0 of the same session, and to
- re-issue the initial query q_0 of this session and look up the set R of ranks of the documents in D (especially of document d) in the corresponding list of retrieved documents.

Again two simplifying assumptions are made here: (a') that every document clicked in a session is relevant for the initial query and (b') that every document relevant for a query q is ranked highly enough in the result list of at least one of the queries in the session starting with the query q such that it is clicked by the user in this session. Assumption (a') is not much stronger than assumption (a) above. Assumption (b') is still drastically simplifying but much weaker than assumption (b) above.

In this sense the statistics collected over the obtained ranks occurring in the set R, even for a huge set of transaction sessions, can never give a perfect picture for the rank distribution but they are getting much closer to anything that could ever be obtained with an automated large-scale method building on assumptions (a) and (b).

Note that such statistics actually approximate the probability of the rank of a document given that it is relevant. Under the assumptions that *(i)* each document is *a priori* equally likely to be relevant and *(ii)* for each possible document every rank is *a priori* equally likely[1], the probability $p(d\,\text{rlv} \mid \text{rank}(d) = r)$ of a document being relevant given its rank is r is proportional to the probability $p(\text{rank}(d) = r \mid d\,\text{rlv})$ of a document having rank r given that it is relevant, due to Bayes' rule

$$p(d\,\text{rlv} \mid \text{rank}(d) = r) = \frac{p(\text{rank}(d) = r \mid d\,\text{rlv}) \cdot p(d\,\text{rlv})}{p(\text{rank}(d) = r)}\,,$$

making our approach valid under the given assumptions.

[1] *A priori* in both cases means before a query is issued.

1.2. Practical contribution

Two experiments were run on a set of sessions extracted from an AOL transaction log[2], using Google for re-issuing the initial queries of the extracted sessions. These experiments involved two different notions of relevance to be discussed below. The results justify many of the assumptions made in the design of applications related to web search: the distribution(s) obtained are actually strictly monotonically decreasing with increasing rank (with negligible exceptions). Interestingly though, despite the time difference between the recording of the AOL log (in 2006) and our empirical study using Google (in 2008) there is still a quite large probability (about 63%) of a relevant document being found within a relatively narrow rank range (1-120). In contrast to that, to find the most relevant document in substantially more than 65% of the cases, not even the top 300 ranked documents are expected to suffice, *i.e.*, the distribution quickly gets very flat (to be discussed below).

Due to several simplifying assumptions the numbers obtained are not to be taken as ground truth; nevertheless there are very clear tendencies to be observed the practical consequences of which are discussed in Section 6.

Note that more details on the approach described below and on applications of the obtained results will be available in [7].

2. Related work

To our knowledge, the question of how the relevant documents are distributed over the ranks in the lists of retrieved documents has not been addressed in the literature so far.

The rank range of documents users view (in the result list for a single query) has been widely studied, showing that users tend to look only at the first ten results and most of the users (percentages close to 80%) tend to not look deeper than two result pages, clicking hardly more than 2 documents per query. The reader is referred to, *e.g.*, [3, 4, 6, 11].

Technically, we made use of previous studies on session identification, see, *e.g.*, [2, 5, 8, 2, 11]. Note that the concrete definition of *session* and thus the methods chosen for session identification necessarily depend on the purpose of the session identification. This results in various techniques reported in the literature a detailed discussion of which is beyond the scope of this paper.

Most common methods compute the difference between the time stamps of two adjacent log records and consider patterns of how queries are changed (called *search patterns*, cf. [2]); some take term overlaps between the corresponding queries into account.

[2] This log was downloaded from http://gregsadetsky.com/aol-data/ and used in [10].

For studies on the influence of search patterns, in combination with the difference of time stamps, the reader is referred to [2]. The system described therein relies on a probabilistic model for session identification based on search patterns. Every pair of adjacent transaction log records is labeled with a search pattern (*e.g.* "generalization", "specialization", "reformulation", etc.). Every search pattern is assigned a probability of the record pair belonging to the same session.

A similar approach based on query content is reported in [9]. Here shifts between sessions are detected with the help of neural networks, also based on time stamp differences and search patterns.

Both these methods involve extensive model training and are computationally expensive. In contrast, a very simple method reported in [5] relies just on the comparison of the query terms in adjacent records. If two adjacent records r_1 and r_2 are generated by requests from the same computer with identical cookies, then r_1 and r_2 are considered to be in the same session if and only if they have at least one term in common. This method is reported to outperform those based on differences in time stamps.

3. Data and preprocessing

For transaction log analysis a log from AOL (records collected from 03/01/2006 until 05/31/2006; 21,011,340 different queries in 36,389,567 transactions from about 650,000 users, see http://gregsadetsky.com/aol-data/) was used. Every transaction record contained the following pieces of information: User ID, query, time stamp, rank of clicked document (if any), clicked URL domain (if any).

In a first preprocessing step, we deleted duplicates and filtered out records for which the query consisted only of a single character and records that consisted only of URLs (records that contained both URLs and non-URL terms were kept). The second preprocessing step was session identification, *i.e.*, partitioning the log into sessions that reflect a sequence of transactions corresponding to the pursuit of one specific information need.

Since several methods for session identification are reported in the literature, we first ran a series of informal experiments to determine which one to use on the resulting datasets. It turned out that for our purpose the most simple method reported, cf. [5], performed just as well as other standard methods. This simple method—which we then used for session identification—defines a session as a maximal sequence of adjacent transaction records by the same user such that every two adjacent queries in this sequence have at least one term in common. However, some tolerance measures were employed to deal with

- spelling mistakes, *e.g.*, in case one record contains the query term *website* and an adjacent record contains the query term *websit* or *wensite*,
- term splits, *e.g.*, in case one record contains the query term *web site* and an adjacent record contains the query term *website*.

To correct spelling mistakes we considered two terms equal if they had a small relative Damerau-Levenshtein distance (in our experiments we set the threshold at 0.25). The relative Damerau-Levenshtein distance [12] between two strings t_1 and t_2 is defined by the number of character insertions, character deletions, character replacements and swaps of adjacent characters that are necessary to obtain t_2 from t_1, divided by the length of the longer one of the two terms.

To deal with term splits, we checked whether combinations of two or three terms in one record occur as a query term in the other record (again with a tolerance in the relative Damerau-Levenshtein distance). In the positive case, the records were considered to have a term in common.

4. Experiment variants

Assumption (a') can be considered in two variants. During a session a user might in general click on more than one listed result. The variant of assumption (a') as discussed in Section 1.1 would consider all these documents relevant for the initial query of the session—the interpretation is that they should be relevant because the user found their snippets offered in the result list relevant enough to click them. A second interpretation might be that the documents clicked before the final transaction in a session are *not* relevant, since the user did not find them satisfactory enough to end the search. With this interpretation, only the last document viewed by the user in a single session is considered relevant for the initial query of the session. These two variants of assumption (a') yield two experiments.

- *AllRel:* In this experiment all documents viewed in a single session are defined relevant for this session.
- *LastRel:* In this experiment only the document viewed after the last query posed in a single session is defined relevant for this session (if there *is* such a document).

After preprocessing we randomly sampled 300,000 sessions in which at least one click occurred, and then picked those that contained relevant documents according to the respective notion of relevance chosen. For each such session, the initial query q_0 was re-issued via the Google API (a special API set up for research programs). In the list L_{q_0} of retrieved results, for every relevant document URL u (for *the* relevant document URL in case of *LastRel*), the aim was to determine the minimal rank of this URL in the list. These ranks, collected over all sessions, would yield our statistics.

However, since the log did not provide the full URLs clicked by the user, but only the domains, we used a two-step approach for every domain name found relevant in a session. Note that for a given domain name *dom* that we associate with a relevant document, we would like to find a full URL u—extending *dom*—that was presumably actually clicked by the user.

First, we issued the query q in the record in which the click for the domain name *dom* occurred (the domain name was clicked since otherwise it would not have been accounted for as relevant). In the result list L_q for q, we looked up the first URL that contained the domain name *dom*. If no such URL occurred below rank 300 of L_q, we defined the URL u to look for by $u =$*dom*; otherwise we took the first URL in L_q containing the domain name *dom* as our relevant URL u.

Second, for our statistics, we counted the rank of the first occurrence of the URL u in the result list L_{q_0} retrieved by Google for the initial query q_0.

5. Results

Out of the 300,000 sessions with clicks that we sampled at random, we extracted all ⟨initial query, relevant document⟩ pairs with the following properties.

- For the *AllRel* experiment, a pair was extracted if the corresponding document was clicked in the session. This resulted in 750,151 such pairs (on average $2.501 \approx 750,151/300,000$ clicks per session).
- For the *LastRel* experiment, a pair was extracted if the corresponding document was clicked in the session *and* this click occurred in the last record of the session. This resulted in 272,377 such pairs (27,623 of the 300,000 sessions with clicks did not have a click in the *last* record).

Table 1 shows the probability distribution (as a percentage) of relevant documents ranked 1-N where $N \in \{20, 120, 300\}$. Note that some relevant documents are not accounted for in this distribution since they occur with a rank > 300 or they do not occur in the result list at all (*e.g.*, because they are no longer indexed by Google).

The corresponding graphs are shown in Figures 1 and 3. These graphs show a *cumulative* probability distribution, *i.e.*, the percentage given for a rank r is the probability that a document is relevant given its rank is at most r. This is obtained by summing up, for all $r' \leq r$, the probability that a document is relevant given its rank is r'.

Note that there is a quite high chance (about 63% for *LastRel*) to find *the* relevant document within a relatively narrow rank range (1-120). In contrast to that, to get a chance substantially higher than 65% (for *LastRel*), not even the top 300 ranked documents are expected to suffice, *i.e.*, the distribution quickly gets very flat. The non-cumulative distributions depicted in Figures 2 and 4 illustrate better how insignificant the probability values for single ranks become after a certain rank.[3]

N	*AllRel*	*LastRel*
≤ 20	43.65%	57.02%
≤ 120	51.06%	62.78%
≤ 300	53.95%	64.93%
> 300	47.05%	35.07%

Table 1. Probability of a document being relevant if its rank is N.

6. Discussion

To interpret our results, the effect of the various assumptions made should be taken into account. The assumptions used to apply Bayes' formula are simplifying, but if we are interested in the shape of the graphs in Figures 3 and 4 rather than the exact values, *i.e.*, when focusing on tendencies, these assumptions are unlikely to be of big impact.

Presumably stronger are the assumptions (a') and (b') which mean that the clicked documents we counted are all relevant and that no other documents are relevant. Since (b') seems more unrealistic than (a'), we assume that the probabilities we obtained up to rank 300 actually underestimate the true ones. Nevertheless, there is no obvious reason to assume that the shapes of the true curves are much different from those we obtain, they might just be shifted in fact.

The log provided only the domains of clicked URLs; to estimate the effect of this it would be reasonable to design further experiments. This is beyond the scope of this paper.

Finally, our distribution may underestimate the true values due to the data used in our 2008 Google experiment being obtained from a 2006 AOL log. It is conceivable that many documents were not found in Google's top 300 hits because they are no longer indexed or have a new URL. Especially, for many sessions new relevant documents are likely to have been added to the index meanwhile.

Despite the fact that the exact values of our distributions are not reliable, our graphs still show important tendencies. As expected, the probability of being relevant decreases with increasing rank. However, it is of practical interest that there is a reasonably small rank number such that relevant documents can be found up to this rank with a reasonably high probability. This to a certain extent justifies re-ranking

[3] Figures 2 and 4 show a nonmonotonicity between ranks 1 to 3. We assume it is related to the fact that we only get domain names from the AOL log, but we have not studied this properly yet.

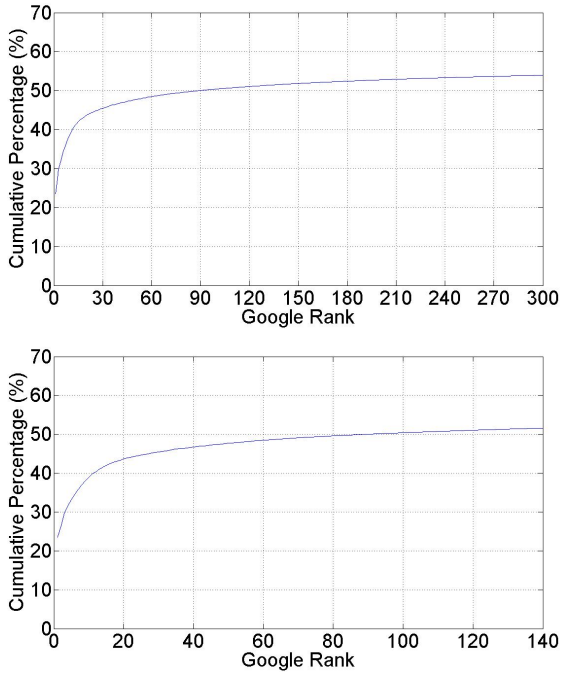

Figure 1. Cumulative probability of a document being relevant given its rank (*AllRel*). Ranks 1-300 and zoom-in for ranks 1-140.

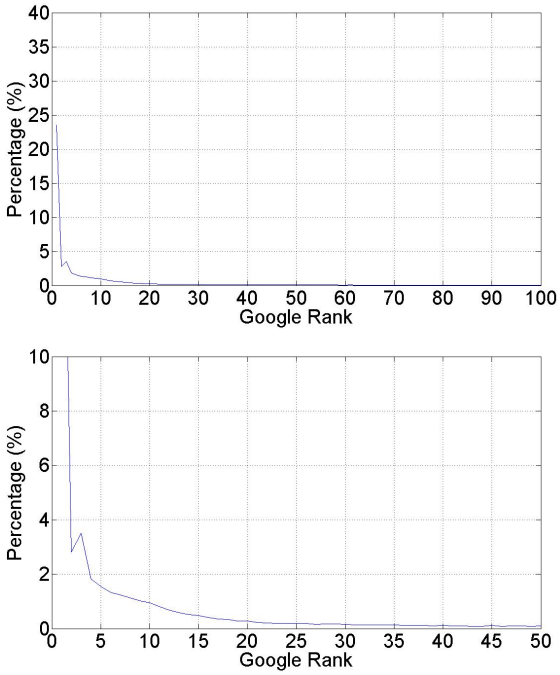

Figure 2. Probability of a document being relevant given its rank (*AllRel*). Ranks 1-100 and zoom-in for ranks 1-50.

methods *etc.* that determine potentially relevant documents based on rank. (If the distribution became flat more quickly such methods would be inappropriate—the rank would give no information about relevance.)

In contrast to this, at an early point our graphs become very flat with a huge portion of the cumulative probability still to be accounted for. Hence increasing the rank range helps only initially to find relevant documents—many of them seem to be hidden very deep in the retrieved list. This should be taken into consideration in practical applications—and it should motivate further research on *(i)* when the distribution actually becomes "too flat" and *(ii)* how high the chance of relevant documents occurring beyond that bound is.

Acknowledgments

We thank Shane Bergsma, Christopher Pinchak (both University of Alberta, Edmonton), and Joel Martin (National Research Council, Ottawa) for their assistance in early stages of this work.

We also gratefully acknowledge support by Google Inc. and the Alberta Ingenuity Centre for Machine Learning.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] D. He, A. Göker, and D. J. Harper. Combining evidence for automatic web session identification. *Information Processing & Management*, 38(5):727–742, 2002.

[3] B. J. Jansen and A. Spink. An analysis of web documents retrieved and viewed. In *International Conference on Internet Computing*, pages 65–69, 2003.

[4] B. J. Jansen and A. Spink. How are we searching the world wide web?: a comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248–263, 2006.

[5] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on web search engines: Research articles. *Journal of the American Society of Information Science and Technology*, 58(6):862–871, 2007.

[6] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36(2):207–227, 2000.

[7] S. Jiang. *Searching for Queries to Improve Document Retrieval in Web Search*. M.Sc. thesis, Department of Computing Science, University of Alberta, to appear.

[8] A. L. Montgomery and C. Faloutsos. Identifying web browsing trends and patterns. *Computer*, 34(7):94–95, 2001.

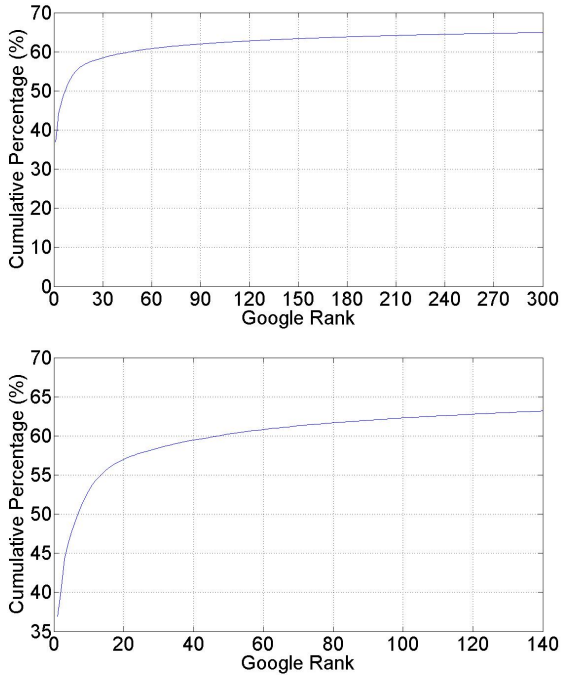

Figure 3. Cumulative probability of a document being relevant given its rank (*LastRel*). Ranks 1-300 and zoom-in for ranks 1-140.

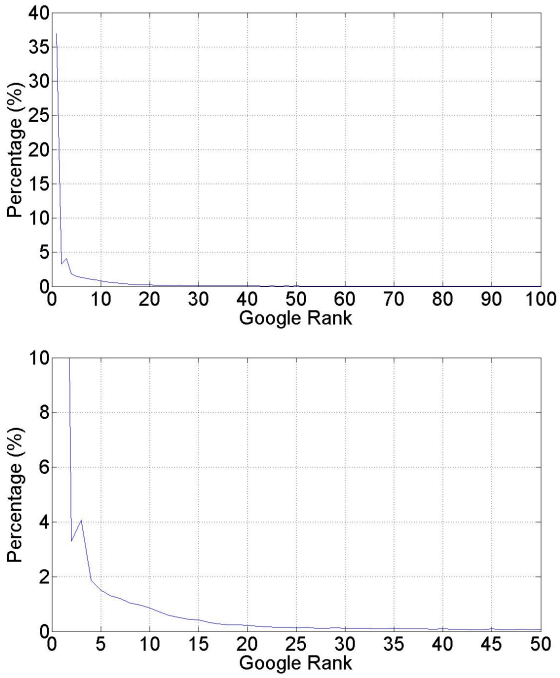

Figure 4. Probability of a document being relevant given its rank (*LastRel*). Ranks 1-100 and zoom-in for ranks 1-50.

[9] S. Özmutlu and F. Çavdur. Neural network applications for automatic new topic identification. *Online Information Review*, 29(1):34–53, 2005.

[10] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proceedings of the International Conference on Scalable Information Systems*, 2006.

[11] C. Silverstein, M. R. Henzinger, H. Marais, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[12] R. A. Wagner and R. Lowrance. An extension of the string-to-string correction problem. *Journal of the ACM*, 22(2):177–183, 1975.