Evaluating the Interest of Revamping Past Search Results

Claudio Orlando Gutierrez Soto, Gilles Hubert

Evaluating the Interest of Revamping Past Search Results

Claudio Gutiérrez-Soto[1,2] and Gilles Hubert[1]

[1] Université de Toulouse, IRIT UMR 5505 CNRS,
118 route de Narbonne, F-31062 Toulouse cedex 9
{Claudio-Orlando.Gutierrez-Soto,Gilles.Hubert}@irit.fr
[2] Departamento de Sistemas de Información,
Universidad del Bío-Bío, Chile

Abstract. In this paper we present two contributions: a method to construct simulated document collections suitable for information retrieval evaluation as well as an approach of information retrieval using past queries and based on result combination. Exponential and Zipf distribution as well as Bradford's law are applied to construct simulated document collections suitable for information retrieval evaluation. Experiments comparing a traditional retrieval approach with our approach based on past queries using past queries show encouraging improvements using our approach.

1 Introduction

Information retrieval (IR) is finding information (usually text) that satisfies an information need from within large collections [12]. An information retrieval system (IRS) represents and stores large amounts of information in order to facilitate and accelerate determination of information estimated relevant to a user's query. A common IRS relies on two main processes: indexing and matching. Indexing intends to construct comparable representations of documents and queries. Matching intends to estimate the extend to which a given document is relevant to a query, usually represented via a score. Documents are then returned to the user in the form of list of results ranked by decreasing score.

IR systems emerged in the late 1940s but improvements really appeared from the late 1950s. Improvements first concerned indexing, ranking functions, weighting schemes, relevance feedback and then models [17]. IR improvements are highly related to evaluation of IRS dating back from the late 1950s [5]. The IR community notably benefited from TREC evaluation campaigns and workshops [21]. These have been offering researchers means to measure system effectiveness and compare approaches.

The literature of IR is crammed with different contributions such as indexing approaches, matching functions, formal models and relevance feedback approaches. However, few approaches gain advantage from searches performed in the past by previous users. Past searches constitute though a useful source of information for new users for example. For example, a user searching about a new subject could benefit from past searches led by previous users about the same subject.

The weak interest of IR in past queries may be understandable because of the lack of suitable IR collections. Indeed, most of the existing IR collections are composed of independent queries. These collections are not usable to evaluate approaches based on past queries since they do not gather similar queries for which ground thruth relevance judgments are provided. In addition, elaborating such collections is difficult due to the cost and time needed. An alternative is to simulate such collections. Simulation in IR dates back to at least the 1980s [10].

In this paper, on the one hand we propose an approach of creating simulated IR collections dedicated to evaluating IR approaches based on past searches. We simulated several scenarios under different probability distributions to build the collection of documents and determine the relevant documents given a query. On the other hand, we introduce a first approach of information retrieval using past queries based on result combination. We experimented this approach on different simulated IR collections using the aforementioned approach. We evaluated the effectiveness of our approach and compared it with a traditional retrieval approach.

This paper is organized as follows. In Sect. 2, related work on simulation in IR evaluation and the use of Past Searches in IR is presented. In Sect. 3 we present our approach to create a simulated IR collection described and the approach of retrieval based on past queries is presented in Sect. 4. Experiments are detailed in Sect. 5. Finally, conclusion and future work are presented in Sect. 6.

2 Related Work

2.1 Simulation in IR Evaluation

The use of simulation in IR is not recent, several approaches have been presented since 1980s. Simulation in IR can be seen as a method where a large collection of queries together with their judgments can be obtained without user interaction [10]. Simulation have been used in different contexts in IR. In [20], an algorithm was developed with the purpose to simulate relevance judgments of users. Here, simulated precision is compared with real precision. In [1], queries to find known-item are simulated. The authors proposed a model to build simulated topics that are comparable to real topics. Works dedicated to simulate the interaction among queries, click log, and preferences of users have been built [6,10]. In addition, today with the exponential growth of the Web, simulation provides interesting approximations of performance on the Web [2,13].

2.2 Past Searches in IR

Many approaches intending to improve results of queries using past searches can be found in the literature. A framework dedicated to improve effectiveness measures such as Average Precision (AP), where a set of documents is assigned to the best system cluster (i.e. best answer given a query) can be found in [3]. Several approaches in IR use past queries for query expansion. For instance, similarity measures are defined in [16] to retrieve past optimal queries that are used to reformulate new queries or to propose the results of past optimal queries. [11] proposed to learn from old queries and their retrieved documents to expand a new submitted query. In [19], historical queries, in particular terms defining queries, are combined to improve average precision for new queries. In [18], feedback information, including clickthrough data and previous queries, are used to improve retrieval effectiveness. Another approach aiming to improve retrieval effectiveness is presented in [4]. A new selection technique using past queries is proposed to estimate utility of available information sources for a given user query. Here, relevance judgments of past queries were not used.

3 Simulating an Information Retrieval Collection

A usual IR collection is composed of three parts: a set of documents,a set of queries and a set of relevance judgments per query (i.e. indications on documents considered relevant or not relevant) [21]. Consequently, our approach aims at defining by simulation. Our method is split in two steps. The first step concerns the creation of terms, documents and queries. The second step involves the simulation of relevance judgments using Bradford's law.

3.1 Creation of Documents and Queries

In this first step, we use an alphabet in to build a set of terms. Each term is composed of letters of this alphabet. This set of terms can be split in subsets called topics in order to represent different subjects. Each letter is chosen using uniform distribution with the purpose to build a term. Thus, each term is unique.

In addition, each document is defined according to all the topics. In order to build a document, topics are selected using either the exponential or Zipf distribution and then terms constituting the document are chosen using uniform distribution. Thus, a document is constructed with terms from one topic mainly but not exclusively.

Past queries are built from documents. To built a past query, a document is chosen under uniform distribution. The terms that constitute the query are chosen from the document under uniform distribution. It is important to emphasize that the intersection among past queries is empty, that is, they have no terms in common. New queries are then built from past queries. For each past query a new query is built, either by changing or adding another term. Thus, the most similar query for the new query is its corresponding past query.

3.2 Simulating Relevant Judgments

In order to simulate the decision given by a user about if a document is relevant or not relevant for a given query, we relied on the zeta distribution. Zeta distribution gives a discrete approximation of Bradford's law [8]. Bradford's law says that among the production of journal papers, there is an heterogeneous number of papers where most relevant papers are in few journals, while a few number of relevant papers are spread on a high quantity of journals. In our case, for a given query, it means that the most relevant documents should be at the top of the list (most relevant papers are in few journal), while a few relevant documents should be spread at down of the list document given a query.

In addition, we assume that for two very similar queries q and q', when a document is relevant for a query, it could be also relevant for the other query. In an intuitive way, there is a subset of common relevant documents for both queries. This does not implies that all relevant documents for query q, are relevant documents for query q'. With the objective to simulate this scenario, we use zeta distribution as follows. We retrieve documents for queries q' and q. Zeta distribution is applied to the set of common documents to the two queries to determine a subset of common relevant documents to q' and q. Eventually, zeta distribution is applied again to retrieved documents of each query q' and q (other relevant documents may be added), preserving the relevant common documents to q and q'. Therefore, the set of relevant documents for q' differs from the set of relevant documents for q.

4 Retrieval Using Past Queries

The basic idea behind of our approach is to incorporate to the system every query with its set of associated documents (query plus its list of documents, which are part of the answer of this query). Thus, the system has not only the set of documents but also the queries executed by user (past queries) with the set of associated documents them. At the beginning just there are documents without the queries, but every time a query is processed by the system, it is aggregated with its documents to the system. When a new query is submitted, first, it is checked and compared with the past queries in the system. When no similar query is found a traditional retrieval can be performed. When similar past queries are found, relevant documents of past queries can be used to build the result for the new query as well as combined with documents retrieved using traditional retrieval. Different strategies combining results from past queries and a traditional result. As first tested approach, we built the result for a new query by adding first the relevant documents from the most similar past query and then documents from a traditional retrieval.

5 Experiments

5.1 Experimental Environment

For this series of experiments, we used the English alphabet in order to build a set of terms. In a general way, the length of a term $|t|$ was between 3 and 7. This length was selected under uniform distribution. The number of terms $|T|$

was 700 for each experiment. We generated documents comprising between 300 to 900 words from a vocabulary composed of between 15 to 30 words for each document [9,15,14]. We defined 7 topics, each topic comprising 100 terms. When a document is built, terms of the other topics are chosen using either exponential distribution or Zipf distribution. Therefore, the most words to compose a document are chosen specific topic. For each experiment we generated five sets of documents comprising 700, 1400, 2100, 2800, and 3500 documents.

On the other hand, we defined 30 queries comprising between 3 to 8 terms for each experiment. The terms of a query were chosen from a particular document. Both terms and documents were chosen using uniform distribution to build the 15 past queries. It is important to mention that intersection among pairs of queries is empty.

In order to simulate judgments of users on documents retrieved given a query q, we implemented the zeta distribution with the purpose to represent the Bradford's law. We applied zeta distribution on the top 30 retrieved documents for each query. Each experiment gathers three different scenarios of zeta distribution for determining relevant documents by varying the parameter s with the values 2, 3, and 4.

5.2 Experimental Results

In this section, three experiments are detailed. We used exponential distribution to build the collection of documents D in the two first experiments and we used Zipf distribution [23] in the third experiment. We computed P@10 (precision at ten retrieved documents) on the results returned by our approach using past results and the traditional cosine retrieval, for each of the thirty queries. Then, we applied the Student's paired sample t-test to test if the difference between the two compared approaches with regards to P@10 was statistically significant.

Experiment 1. Exponential distribution with parameter equal to 1.5 was used to build the dataset D. When using zeta distribution with parameter $s = 2$ for relevance, our approach improved P@10 for 26 over 30 queries on average over the five sets of documents generated (i.e. comprising 700, 1400, 2100, 2800, and 3500 documents). The average P@10 improvement was +35.00 % over all queries. Statistical significance was reached in all cases, the highest p-value for the Student's t-test being 0.00128 for the set of 700 documents. When using zeta distribution with parameter $s = 3$, our approach improved P@10 for 24.4 over 30 queries on average over the five sets of documents. The average P@10 improvement was +33.99 %. Statistical significance was reached in all cases, the highest p-value for the Student's t-test being 0.00056 for 3500 documents. When using zeta distribution with parameter $s = 4$, our approach improved P@10 for 21.4 over 30 queries on average over the five sets of documents. The average P@10 improvement was +38.48 %. Statistical significance was reached in all cases, the highest p-value for the Student's t-test being 0.00732 for 1400 documents.

Experiment 2. Exponential distribution with parameter equal to 1.0 was used to build the dataset D. When using zeta distribution with parameter $s = 2$ for relevance, our approach improved P@10 for 26.2 over 30 queries on average over the five sets of documents. The average P@10 improvement was +31.12 %. Statistical significance was reached in all cases, the highest p-value for the Student's t-test being 0.00002 for 1400 documents. When using zeta distribution with $s = 3$, our approach improved P@10 for 21.6 over 30 queries on average. The average P@10 improvement was +26.10 %. Statistical significance was reached in all cases, the highest p-value for the Student's t-test being 0.01306 for 3500 documents. When using zeta distribution with parameter $s = 4$, our approach improved P@10 for 21.6 over 30 queries on average. The average P@10 improvement was +33.76 %. Statistical significance was reached in all cases, the highest p-value for the Student's t-test being 0.00122 for 2800 documents.

Experiment 3. Zipf distribution with parameter equal to 1.6 was used to build the dataset D. When using zeta distribution with parameter $s = 2$ for relevance, our approach improved P@10 for 24.8 over 30 queries on average over the five sets of documents. The average P@10 improvement was +25.50 %. Statistical significance was reached in all cases, the highest p-value for the Student's t-test being 0.00005 for 700 documents. When using zeta distribution with parameter $s = 3$, our approach improved P@10 for 22.6 over 30 queries on average. The average P@10 improvement was +22.77 %. Statistical significance was reached in all cases, the highest p-value for the Student's t-test being 0.00034 for 1400 documents. When using zeta distribution with parameter $s = 4$, our approach improved P@10 for 24.8 over 30 queries on average. The average P@10 improvement was +27.92 %. Statistical significance was reached in all cases, the highest p-value for the Student's t-test being 0.00031 for 2100 documents.

5.3 Discussion

Due to space limitations, some experiment details were not presented, however additional observations are reported in this section. Using different parameters in zeta distribution ($s = 2, 3$ and 4) for relevance judgments allowed us to analyze how function influences on the average P@10. According to the evaluations we observed that average P@10 decreases for both compared approaches when the parameter s is increases in zeta distribution. In addition, we observed that there is not a radical tendency in the differences of average P@10 between the approach using past queries and the traditional retrieval approach when the number of documents increases.

Summarizing the results reported in this paper, each experiment and each scenario showed that the approach based on past queries always overcomes the traditional approach ($>$ +22.50 %). Statistical significance was always reached since the highest p-value for the Student's t-test was 0.01306.

In addition one should notice that we applied the Zipf distribution with value 1.6 only to simulate distribution frequencies of terms from topics for document creation. It would be interesting to test other Zipf distributions to simulate the distribution of the frequencies of terms [23]. It is for this reason that we used not only the Zipf distribution but also exponential distribution to build collections of documents.

Eventually, our experiments used simulated datasets which allow us to have promising preliminary results. However, we should test other datasets using other distributions such as power law for term selection or using queries generated from different document collections.

6 Conclusion and Future Work

In this paper, we have presented on the one hand an approach of information retrieval using past queries. This approach is based on the reuse, for a new submitted query, of relevant documents retrieved for the most similar past query. On the other hand, due to the lack of available existing IR collections suitable for evaluating this kind of approach, we proposed an approach to creating simulated IR collections. We simulated several scenarios under different probability distributions to build the collection of documents and determine the relevant documents given a query. We experimented our approach of retrieval based on past queries on different simulated IR collections using the aforementioned approach. We evaluated the effectiveness of our approach and compared it with a traditional retrieval approach. Experiments showed encouraging results when evaluating the results for the top ten retrieved documents (P@10).

Future work will be devoted first to define more real evaluation datasets suitable for IR approaches based on past queries, adapting TREC collections for instance. We will also develop other approaches to construct the retrieved documents for a new query from various results of past queries, based on clustering methods [22] and based on diversification methods [7].

References

1. Azzopardi, L., de Rijke, M., Balog, K.: Building simulated queries for known-item topics: an analysis using six european languages. In: Proceedings of the 30th annual international ACM SIGIR, pp. 455–462. ACM, New York (2007)
2. Baeza-Yates, R., Castillo, C., Marin, M., Rodriguez, A.: Crawling a country: better strategies than breadth-first for web page ordering. Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, WWW 2005, pp. 864–872. ACM, New York (2005)
3. Bigot, A., Chrisment, C., Dkaki, T., Hubert, G., Mothe, J.: Fusing different information retrieval systems according to query-topics: a study based on correlation in information retrieval systems and trec topics. Inf. Retr. 14(6), 617–648 (2011)
4. Cetintas, S., Si, L., Yuan, H.: Using past queries for resource selection in distributed information retrieval. Tech. Rep. 1743, Department of Computer Science, Purdue University (2011), `http://docs.lib.purdue.edu/cstech/1743`
5. Cleverdon, C.W.: The evaluation of systems used in information retrieval (1958: Washington). In: Proceedings of the International Conference on Scientific Information - Two Volumes, pp. 687–698 (1959)

6. Dang, V., Croft, B.W.: Query reformulation using anchor text. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM 2010, pp. 41–50. ACM, New York (2010)
7. Drosou, M., Pitoura, E.: Search result diversification. SIGMOD Rec. 39(1), 41–47 (2010)
8. Garfield, E.: Bradford's Law and Related Statistical Patterns. Essays of an Information Scientist 4(19), 476–483 (1980), `http://www.garfield.library.upenn.edu/essays/v4p476y1979-80.pdf`
9. Heaps, H.S.: Information Retrieval: Computational and Theoretical Aspects. Academic Press, Inc., Orlando (1978)
10. Huurnink, B., Hofmann, K., de Rijke, M., Bron, M.: Validating query simulators: An experiment using commercial searches and purchases. In: Agosti, M., Ferro, N., Peters, C., de Rijke, M., Smeaton, A. (eds.) CLEF 2010. LNCS, vol. 6360, pp. 40–51. Springer, Heidelberg (2010)
11. Klink, S.: Improving document transformation techniques with collaborative learned term-based concepts. In: Dengel, A.R., Junker, M., Weisbecker, A. (eds.) Adaptive READ Research Project. LNCS, vol. 2956, pp. 281–305. Springer, Heidelberg (2004)
12. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (July 2008)
13. Marin, M., Gil-Costa, V., Bonacic, C., Baeza-Yates, R., Scherson, I.D.: Sync/async parallel search for the efficient design and construction of web search engines. Parallel Comput. 36(4), 153–168 (2010)
14. Silva de Moura, E., Navarro, G., Ziviani, N., Baeza-Yates, R.: Fast and flexible word searching on compressed text. ACM Trans. Inf. Syst. 18(2), 113–139 (2000)
15. Navarro, G., De Moura, E.S., Neubert, M., Ziviani, N., Baeza-Yates, R.: Adding compression to block addressing inverted indexes. Inf. Retr. 3(1), 49–77 (2000)
16. Raghavan, V.V., Sever, H.: On the reuse of past optimal queries. In: Proceedings of the 18th Annual International ACM SIGIR Conference, pp. 344–350. ACM, New York (1995)
17. Sanderson, M., Croft, W.: The history of information retrieval research. Proceedings of the IEEE 100(Special Centennial Issue), 1444–1451 (2012)
18. Shen, X., Tan, B., Zhai, C.: Context-sensitive information retrieval using implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference, pp. 43–50. ACM, New York (2005)
19. Shen, X., Zhai, C.X.: Exploiting query history for document ranking in interactive information retrieval. In: Proceedings of the 26th Annual International ACM SIGIR Conference, pp. 377–378. ACM, New York (2003)
20. Tague, J.M., Nelson, M.J.: Simulation of user judgments in bibliographic retrieval systems. In: Proceedings of the 4th Annual International ACM SIGIR Conference, pp. 66–71. ACM, New York (1981)
21. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge (2005)
22. Xu, R., Wunsch, D.I.: Survey of clustering algorithms. IEEE Transactions on Neural Networks 16(3), 645–678 (2005)
23. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Addison-Wesley, Reading (1949)