Gilles Hubert, Guillaume Cabanac, Christian Sallaberry, Damien Palacio. Query Operators Shown Beneficial for Improving Search Results. In Theory and Practice of Digital Libraries (TPDL), Berlin, Germany, 26/09/11-28/09/11, Springer, LNCS, Sept. 2011 (to appear).

Query Operators Shown Beneficial for Improving Search Results

Gilles Hubert[1], Guillaume Cabanac[1],
Christian Sallaberry[2], and Damien Palacio[2]

[1] Université de Toulouse, IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9

[2] Université de Pau et des Pays de l'Adour, LIUPPA ÉA 3000
Avenue de l'Université, BP 1155, F-64013 Pau cedex

Abstract. Search engines allow users to retrieve documents with respect to a given query. These provide advanced search options, such as query operators (e.g., `+term`, `term^10`). Previous work studied how query operators are employed by end-users. In this paper, we study the extent to which using query operators may lead to improved results, regardless of specific users. We hypothesize that the proper use of query operators improves search results. To validate this hypothesis, we present a methodology relying on standard IR test collections. We applied this methodology to TREC-7 and TREC-8 test collections with five IR models implemented in the Terrier search engine. Experiments show that queries enriched with operators give an improvement in effectiveness up to 35.1% over regular queries. This result suggests that end-users would benefit from using operators more often.

Keywords: Information Retrieval, Search Engine, Query Operators, Effectiveness

1 Introduction

Digital Libraries (DL), such as Europeana [15], aim to collect and give access to huge amounts of multimedia documents. People may either browse these repositories or retrieve documents matching their interests, thanks to a search engine. In the latter case, users have to translate their mental information needs into a sequence of terms, called a query. For instance, a scientist looking for research projects funded in the DL domain may issue the following query: [`digital library research project funding`].

For reducing the mismatch between the user's cognitive model of the need and the produced query, search engines offer query operators [17], as a way to specify the role of (group of) terms. These comprise boolean operators (e.g., `AND`, `OR`, `NOT`), expressions surrounded by quotation marks (e.g., `"digital libraries"`), proximity operators (e.g., `NEAR`), boosting operators, and so on. In the previous example, the scientist may expect better results when wording his/her query as: [`"digital library" +research project funding^4`]. This refined query

better conveys the intent of the user, since the search engine is instructed that the research aspect is mandatory, and the funding aspect should be considered as a strong evidence for ranking documents in the result list.

Search engines foster the use of query operators to improve retrieval quality, as reported in [4]. The real use of query operators was studied in [17,18,4,23,1] by analyzing logs of popular search engines, such as Altavista, Excite, Google, MSN Search, and Yahoo!. Researchers found that queries with operators constitute up to 20% of all submitted queries. This recognized use raised questions about the effects of operators on search accuracy. While operators are expected to improve search engine effectiveness, as underlined by White and Morris [23], experiments by Eastman and Jansen [4] showed that the outcome may be 'relatively small.' The reasons behind this poor improvement may be related to various factors ranging from users to Information Retrieval (IR) models implemented in search engines. Although previous work was devoted to understanding the 'user' factor [17,18,11,4,5,1], we failed to find any research investigating the 'system' parameter. We therefore analyze in this paper the potential of effectiveness improvement yielded by operators, regardless of specific users. We address hypothesis $\mathcal{H}$: *the proper use of query operators improves search results.*

The paper is organized as follows. In Sect. 2, we review the literature devoted to query operators as featured by search engines. We stress that studies to date did not measure the potential of effectiveness improvement when properly using query operators. In Sect. 3, we present the proposed evaluation methodology involving experiments with standard IR test collections. In Sect. 4, we report the results of the experiments that we conducted with the TREC-7 [20] and TREC-8 [21] test collections. These validate $\mathcal{H}$: *the proper use of query operators does improve search results.* We conclude the paper in Sect. 5, and give insights into future work.

2 Related Work: Operators in Search Queries

Studies of operators in queries submitted to search engines fall into the two categories discussed in the following sections.

2.1 Usage of Query Operators

Several studies reported the use of query operators for common search engines. Analyzing various query logs with different characteristics (number of queries, users, crawling timespan), researchers found the following proportions of queries with operators — due to space limitation, we emphasize on most recent research:

- For Altavista, Silverstein et al. [17] found 20.4% in 1999.
- For Excite, Jansen et al. [10] found 24.1% in 2000, while Spink et al. [18] found 14.5% in 2001.
- For Google, MSN Search, and Yahoo! altogether, White and Morris [23] found 1.12% in 2007. Notice, however, that this study was limited to four operators (i.e., `+`, `-`, `"..."`, and `site:`).

In addition to these quantitative studies, other research was concerned with qualitative analysis related to users. Hölscher and Strube [8], as well as Lucas and Topi [11], found that expert users recourse to query operators more frequently than the average user. Using query operators is a trait that one would expect from expert searchers, according to White and Morris [23]. Jansen et al. [10] point that average users "are certainly not comfortable with Boolean operators and other advanced means of searching."

When present in queries, operators are used in a "semantically appropriate manner," according to Eastman and Jansen [5]. Users tend to use more operators when facing complex information needs and having difficulty in finding information [1]. Overall, query operators were found to be used more in dedicated search engines (e.g., online DL catalogues) than in web search engines [9].

2.2 Benefits Brought by Using Query Operators

Beyond measuring the proportion of queries with operators in search engine query logs, a few studies investigated their effects on retrieval effectiveness. Eastman and Jansen [4] stated in 2003 that only few studies compared retrieval results using query variants (i.e., with and without operators). Since then, White and Morris confirmed that observation: query operators "have generally been overlooked by the research community in attempts to improve the quality of search results" [23].

One prominent study measuring the effect of query operators on result accuracy was conducted by Eastman and Jansen [4]. Their experiment involved two sets of queries (A and B). Set A contained 100 original queries with operators (`AND`, `OR`, `MUST APPEAR`, and `PHRASE`), which were extracted from Excite logs. Set B contained the 100 original queries with all operators removed. Finally, queries from sets A and B were submitted to three search engines: AOL, Google, and MSN Search. The top 10 documents retrieved by each search engine, for each query, were judged by 4 experts on a 4-point scale. Documents marked with an average score of 3 or higher were considered as relevant to the query. Among other measures, averaged 'relative precision' P@10 (i.e., number of relevant documents in the top 10) was calculated for sets A and B. The researchers showed that the recourse to query operators (set A) did not yield statistically significant improvement over operator-free queries (set B). They concluded that "the use of most operators had no significant effect on ... relative precision."

Eastman and Jansen's [4] conclusions may lead searchers to get rid of query operators. Nevertheless, we wonder why this study focused on queries with operators in the first place, since these only represent up to 20% of all submitted queries. In addition, these are known to be more complex than average queries [1]. That is the reason why we ask ourselves, in this paper, whether their conclusions still hold for the 80% remaining queries that users formulate without operators. To do that, we intend to evaluate the improvement in effectiveness yielded by refining regular queries with operators. The next section introduces the methodology that we designed for that purpose.

3 Methodology: Assessing the Effects of Query Operators

We designed an evaluation methodology to test $\mathcal{H}$, that is: *the proper use of query operators improves search results*. Two research questions arose when trying to validate this hypothesis:

- Q_1. What is the maximum gain in effectiveness that one can expect by enriching a query with operators only (no term modification or addition)?
- Q_2. Do users succeed in formulating queries with operators, so that these lead to a significant gain in effectiveness?

If we notice no possible gain when using query operators (Q_1), we obviously cannot expect users to get better results when having recourse to them (Q_2). Hence, the answer to Q_2 depends on the answer to Q_1. We thus focus on answering Q_1 in this paper. The proposed methodology is illustrated in Fig. 1. It involves the four stages that we detail in the following sections.

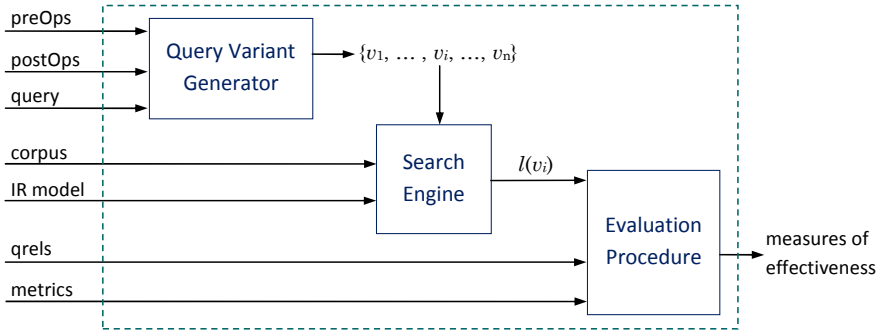

Fig. 1. Illustration of the methodology for assessing the effects of query operators

3.1 Selection of an IR Test Collection

An IR Test Collection allows researchers to run experiments for evaluating their search engines [16]. It is comprised of four components:

1. The *corpus* is a set of documents to be indexed by the search engine.
2. A set of n *topics* represents user information needs. Each topic may be worded as:
   - A *title*: a sequence of two or three terms (in general) that would be submitted as a query to a search engine by an average user.
   - A *description*: a few sentences describing the user's information needs in plain text.
   - A *narrative*: a longer text than the description, which gives insights into the intent of the user, and unambiguously states what information is relevant or irrelevant for the searcher.

3. The *qrels* a.k.a. query relevance judgments, or gold standard. Usually, experts identify and mark documents according to their relevance to each topic. These marks/ratings may be binary (i.e., nonrelevant vs. relevant) or gradual (e.g., ranging from nonrelevant to relevant on a 5-point scale).
4. The *metrics* allow the measurement of the accuracy of the results retrieved by a search engine. As detailed in [2], common metrics are Precision, Recall, F1, Average Precision, and so on. Taking as input (i) the given topic and (ii) the result list produced by the search engine, a metrics computes a numerical value representing the effectiveness of the search engine in retrieving relevant documents to the user.

Several IR Test Collections were produced by IR initiatives and then released to the community for research purposes. As the prominent evaluation initiative in IR, TREC [22] has been providing many collections [7] since 1992. Note that there exists other initiatives with similar purposes (e.g., CLEF, NTCIR).

3.2 Generation of Query Variants with Tested Operators

We intend to check that any given user's query can be rewritten with operators, such that it leads to more accurate search results. In the remainder of the paper, we call 'preOps' the operators prefixing a query term (e.g., ∅, `+`), and 'postOps' the operators postfixing a query term (e.g., ∅, `^2`, `^10`). For each *topic* in the selected Test Collection, we consider the associated *title* (or *description*, or *narrative*) comprising t terms. Then, a Query Variant Generator is used to generate $\omega = (b \cdot a)^t$ variants with the b given preOps, and the a postOps to be tested. These are denoted $\{v_1, \ldots, v_i, \ldots, v_n\}$ in Fig. 1.

Example 1. Let us consider a *topic* from a Test Collection, with the following *title*: [`1974 Turing award recipient`]. Generating variants for the $b = 2$ preOps $\in \{\varnothing, \texttt{+}\}$, and $a = 3$ postOps $\in \{\varnothing, \texttt{\^{}2}, \texttt{\^{}10}\}$ leads to $\omega = (2 \cdot 3)^4 = 1{,}296$ query variants illustrated in Table 1.

Table 1. Excerpt of the 1,296 query variant generated with prefix operators $\{\varnothing, \texttt{+}\}$ and postfix operators $\{\varnothing, \texttt{\^{}2}, \texttt{\^{}10}\}$

Variant #	Query variants generated with preOps and postOps			
1	`+1974`	`Turing`	`award`	`recipient`
2	`1974`	`+Turing`	`award`	`recipient`
⋮	⋮	⋮	⋮	⋮
1,295	`+1974^10`	`+Turing^10`	`+award^10`	`+recipient^2`
1,296	`+1974^10`	`+Turing^10`	`+award^10`	`+recipient^10`

3.3 Retrieval with Initial Query and Generated Variants

During this third stage, the documents from the *corpus* are indexed by a Search Engine. Then, it is run according to a given IR model, which governs the way queries are matched with documents. TF·IDF and OkapiBM25 may be cited as examples of prominent IR models. Since detailing how these models operate is beyond the scope of this paper, we refer the reader to [3, chap. 7] for a comprehensive presentation of this topic.

Finally, the initial (operator-free) query q, and the generated ω query variants v_i (with operators) are submitted to the search engine. For query q is retrieved a list of documents $l(q)$. For any query variant v_i is retrieved a list of document $l(v_i)$. Note that all lists of documents are ranked by decreasing RSV (i.e., Retrieval Status Value: the relevance score estimated with respect to the query).

3.4 Data Analysis: Measuring Effectiveness Variations

The fourth stage of the methodology is concerned with the analysis of search results. An Evaluation Procedure applies a metric m to the document list $l(q)$ or $l(v_i)$, and to the *qrels* $j(q)$ associated with the tested initial query q. The value of this metrics $m(l(q), j(q)) \in [0, 1]$, also called 'measure,' represents the extent to which query q yielded relevant results. Similarly, $m(l(v_i), j(q)) \in [0, 1]$ represents the extent to which query variant v_i yielded relevant results. According to these measures, one may report per topic analyses, as well as global analyses.

Per Topic Analysis. For a given evaluation metrics m (e.g., Recall, Average Precision), a given initial query q, and its variants $v_{i \in [\![1,\omega]\!]}$, one gets $\omega + 1$ measures. These data values represent the outcome when applying query operators to initial query q. Among them, the *maximum value* $\max_{i=1}^{\omega} m(l(v_i), j(q))$ is the best performance reachable by using operators properly.

In addition, one may study the distribution of effectiveness data values thanks to the 'boxplot' visualization [19,24]. As shown in Fig. 2, a boxplot (a.k.a. box-and-whisker diagram) summarizes several descriptive statistics. The interquartile range (IQR) spans the lower quartile to the upper quartile. The middle 50% of the ranked data lies in the IQR. It is represented as a *box* (central rectangle), which shows the spread of data values. The median is shown as a *segment* inside the box. This is the middle half of the data values, and allows one to assess the symmetry of the distribution. The *whiskers* extend from the ends of the box to the most distant value lying within $1.5 \times$ IQR. Larger and lower values are considered as outliers; these are plotted with *black circles*.

For a given *topic*, we may finally compare the effectiveness of the initial query q with the effectiveness of the best query variant v_i. Equation (1) computes the percent gain yielded by query operators.

$$g(q, v_i) = \frac{m(l(v_i), j(q))}{m(l(q), j(q))} - 1 \tag{1}$$

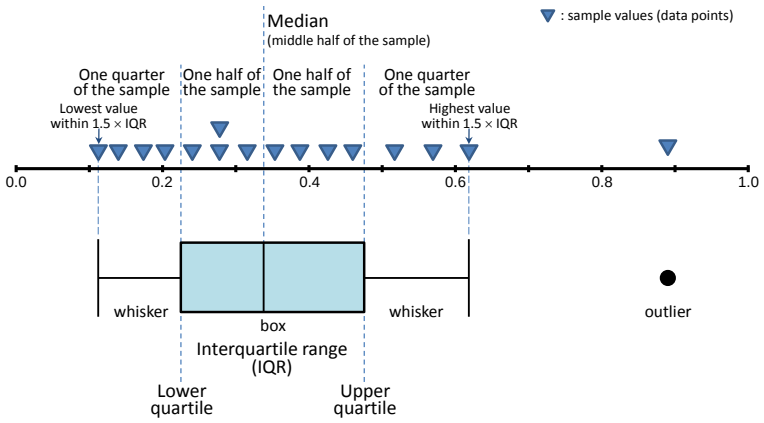

Fig. 2. Example of data values and associated boxplot

Global Analysis. Controlling for 'topic effect' is usually done by evaluating the search engine with n topics, and then averaging out the individual n effectiveness scores (e.g., AP meaning Average Precision) in a global score (e.g., MAP meaning Mean Average Precision). This practice was shown to give stable conclusions with a least $n = 25$ topics, while $n = 50$ is the standard at TREC [2].

The comparison between MAP of initial query q, and MAP of best query variant v_i allows the checking of $\mathcal{H}$. For validating this hypothesis, we must show that (i) v_i is more effective than q, and that (ii) the observed difference is statistically significant with regard to Student's bilateral paired t-test. According to the resulting p-value, the two data samples are said to be statistically different when $p < 0.05$. The interested reader is referred to [16] for in-depth coverage of statistical testing in IR.

4 Experiments and Results

We applied the devised methodology to TREC-7 [20] and TREC-8 [21] standard test collections. They provide a corpus of newspaper articles; this corresponds to the kind of documents that a DL would index (contrary to other TREC test collections providing web documents). Moreover, they provide $n = 100$ topics covering various subjects, allowing us to conduct significance testing.

We tested two query operators: (i) the 'must appear' prefix operator (`+`), as it is considered as popular [8] and "easy to employ and available" [4], as well as (ii) the boosting postfix operator (`^N`), as we found no study to date on this operator. Among search engines supporting these both features (e.g., Lemur [12], Lucene [6], Terrier [13]), we used Terrier version 3.0, which provides several IR models. In order to check the 'model effect' on results, we conducted experiments with the following five models: BM25, DFR_BM25, InL2, PL2, and TF_IDF. We queried Terrier with the *title* part (see Sect. 3.1) of the 100 topics, the 9,953 variants generated for TREC-7, and the 11,203 variants generated for TREC-8.

4.1 Per Topic Analysis

We report in Fig. 3 the per topic effectiveness (AP) of the baseline (initial query q without operators, plotted as red diamonds ◆), as well as the boxplots summarizing the APs of all query variants. We used the default model provided in Terrier [13], namely PL2. Overall, regarding both TREC-7 and TREC-8 results, the baseline AP is highly variable, suggesting that these test collections feature a mixture of easy and hard topics.

Regarding mono-term queries (e.g., Topic 14 for TREC-7 or Topic 3 for TREC-8), we noticed no difference between AP of q and AP of v_i. This result was expected since a mono-term is mandatory *per se*, and boosting has effect with multi-term queries only.

For the remaining queries, the baseline AP generally lies over the median (i.e., segment inside the central rectangle) in Fig. 3. This suggests that most variants led to worse AP than the initial query q did. Nevertheless, there is always a query variant v_i whose AP equals or outperforms the AP of the baseline query q. This observation tends to support the hypothesis $\mathcal{H}$. This improvement does not seem to depend on the AP of the baseline: there is way for improvement for poor baselines, as well as for strong baselines.

4.2 Global Analysis

In addition to per topic analysis, we quantify the observed performance differences thanks to a global analysis covering the n topics of TREC-7 and TREC-8. We measured the MAP of the baseline (queries without operators), and the MAP of the best variant per topic (queries with operators), for each of the five tested IR models. Then, we computed the gain brought by operators (Δ in percent, as reported in Tables 2–3) to validate hypothesis $\mathcal{H}$.

We conducted three experiments: (i) with operators $\{\varnothing, \texttt{+}\}$, (ii) with operators $\{\varnothing, \texttt{\^{}10}, \texttt{\^{}20}, \texttt{\^{}30}, \texttt{\^{}40}, \texttt{\^{}50}\}$ featuring arbitrarily chosen weights, and (iii) a combination of the two.

Table 2. Evaluation results with 'must appear' operator (+)

	TREC-7			TREC-8		
	MAP			MAP		
Model	Baseline	VOP	Δ(%)	Baseline	VOP	Δ(%)
BM25	0.1677	0.1836	9.5**	0.1957	0.2154	10.2*
DFR_BM25	0.1683	0.1843	9.5**	0.1965	0.2162	10.0*
InL2	0.1710	0.1852	8.3**	0.1996	0.2172	8.8*
PL2	0.1554	0.1826	17.5**	0.1840	0.2106	14.5**
TF_IDF	0.1674	0.1833	9.5**	0.1964	0.2158	9.9**

Statistical significance is denoted by '*' for $p < 0.05$ ('**' for $p < 0.01$)

In Table 2, queries with operator (VOP) always overcome the baseline, whatever the IR model. Notice that the differences are statistically significant. In

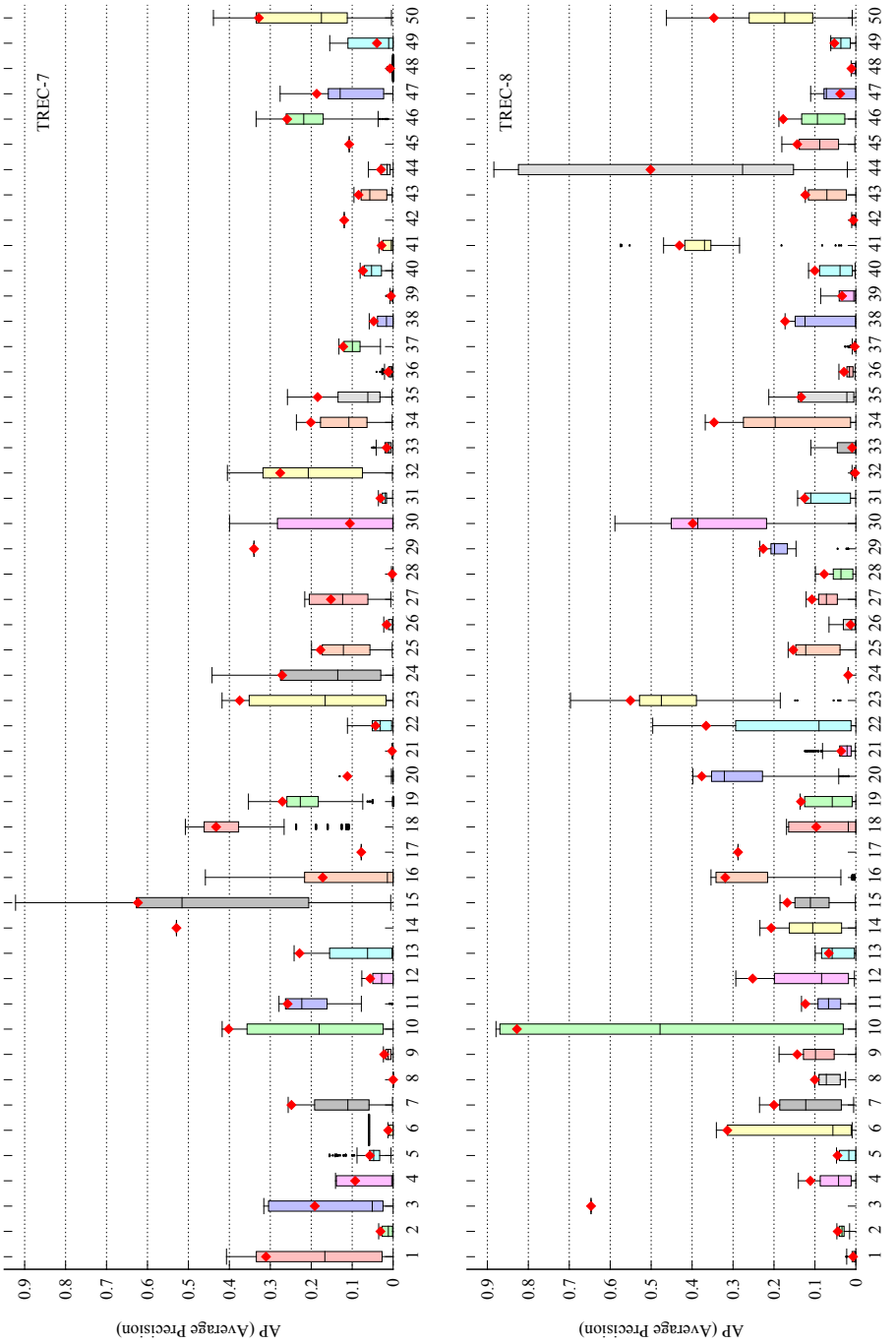

Fig. 3. Average Precision (AP) boxplots showing potential for improvement with query operators on 100 queries (TREC-8 and TREC-7) with model PL2 of Terrier [13]. Diamonds ◆ show the AP of TREC's initial operator-free query

Table 3, queries with operator (VOB) always overcome the baseline, whatever the IR model. Notice that the differences are statistically significant. Overall, the boosting operator yields better results than the 'must appear' operator.

Table 3. Evaluation results with boosting operator (`^N`)

	TREC-7			TREC-8		
	MAP			MAP		
Model	Baseline	VOB	$\Delta(\%)$	Baseline	VOB	$\Delta(\%)$
BM25	0.1677	0.2027	20.9^{**}	0.1957	0.2312	18.1^{**}
DFR_BM25	0.1683	0.2034	20.9^{**}	0.1965	0.2316	17.9^{**}
InL2	0.1710	0.2059	20.4^{**}	0.1996	0.2352	17.8^{**}
PL2	0.1554	0.1926	23.9^{**}	0.1840	0.2173	18.1^{**}
TF_IDF	0.1674	0.2026	21.0^{**}	0.1964	0.2312	17.7^{**}

Statistical significance is denoted by '*' for $p < 0.05$ ('**' for $p < 0.01$)

In Table 4, queries with operator (VOPB) always overcome the baseline, whatever the IR model. Notice that the differences are statistically significant. Overall, the combination of 'must appear' and boosting operators yields best results, up to a 35.1% improvement. This is a material improvement, which would be even larger if we had got rid of mono-term queries, since they do not benefit from any operator for sure.

Table 4. Evaluation results with boost and 'must appear' operators (`+` and `^N`)

	TREC-7			TREC-8		
	MAP			MAP		
Model	Baseline	VOPB	$\Delta(\%)$	Baseline	VOPB	$\Delta(\%)$
BM25	0.1677	0.2132	27.1^{**}	0.1957	0.2381	21.7^{**}
DFR_BM25	0.1683	0.2133	26.7^{**}	0.1965	0.2387	21.5^{**}
InL2	0.1710	0.2144	25.4^{**}	0.1996	0.2407	20.6^{**}
PL2	0.1554	0.2099	35.1^{**}	0.1840	0.2288	24.3^{**}
TF_IDF	0.1674	0.2131	27.3^{**}	0.1964	0.2383	21.3^{**}

Statistical significance is denoted by '*' for $p < 0.05$ ('**' for $p < 0.01$)

4.3 Discussion of Results

In our experiments, Terrier [13] was considered as a black box. To our knowledge, most IR models do not specify how to handle mandatory terms, as well as boosted terms. Terrier, however, implements this feature, which suggests that it performs a specific computation.

Regarding the reported results for the boosting operator (Tables 3–4), we selected boosting weights arbitrarily. Other values may have given different results. We leave to future work the study of boosting weights on effectiveness.

5 Conclusion and Future Work

Previous work considered the use of query operators in common search engines. Eastman and Jansen [4] notably studied whether queries with operators yield similar effectiveness with respect to counterparts without operators. They reported a limited improvement due to operators, which questions the return on investment that users may grant to query operators. We wondered if this poor effect was due to users or search engines.

In this paper, we considered the majority (80%) of queries submitted to search engines: those without operators. We stated hypothesis $\mathcal{H}$: *the proper use of query operators improves search results*. We designed a methodology to validate $\mathcal{H}$ through the use of standard IR test collections, and the generation of query variants with operators. We applied this methodology using TREC-7 and TREC-8 test collections. Experiments showed that TREC's initial query can always be improved by refining it with 'must appear' and boosting operators. The observed gain — up to 35.1% — is statistically significant, whatever the tested IR model and collection. This suggests that, when properly used, users benefit from refining queries with such operators. Indeed, query operators convey information instructing the search engine about requirements and preferences (as expressed in the *narrative* part of topics) that would remain implicit otherwise.

Directions for future work include, in the short term, experimenting our methodology in various contexts (e.g., additional IR collections, IR models, query operators). In the medium term, we plan to address Q_2 stated in Sect. 3: Do users succeed in formulating queries with operators, so that these lead to a significant gain in effectiveness? In addition, we should study other factors involved when retrieving information [11], such as the number of terms used, and the selection of terms. In the long term, we may study operator use and effects for retrieval involving more than the topical dimension of information. This is notably the case of geographic IR [14] involving spatial and temporal dimensions in addition to the topical dimension of information.

References

1. Aula, A., Khan, R.M., Guan, Z.: How does search behavior change as search becomes more difficult? In: CHI'10: Proceedings of the 28th international conference on Human factors in computing systems. pp. 35–44. ACM, New York, NY, USA (2010)
2. Buckley, C., Voorhees, E.M.: Retrieval System Evaluation. In: Voorhees and Harman [22], chap. 3, pp. 53–75
3. Croft, W.B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Addison-Wesley (Feb 2010)
4. Eastman, C.M., Jansen, B.J.: Coverage, relevance, and ranking: The impact of query operators on web search engine results. ACM Trans. Inf. Syst. 21(4), 383–411 (Oct 2003)

5. Eastman, C.M., Jansen, B.J.: The appropriate (and inappropriate) use of query operators and their effect on web search results. Proceedings of the American Society for Information Science and Technology 41(1), 274–279 (2004)
6. Gospodnetić, O., Hatcher, E.: Lucene in Action. Manning Publications (2005)
7. Harman, D.K.: The TREC Test Collections. In: Voorhees and Harman [22], chap. 2, pp. 21–53
8. Hölscher, C., Strube, G.: Web search behavior of internet experts and newbies. Comput. Netw. 33, 337–346 (Jun 2000)
9. Jansen, B.J., Pooch, U.: A review of web searching studies and a framework for future research. J. Am. Soc. Inf. Sci. Technol. 52(3), 235–246 (Feb 2001)
10. Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. Inf. Process. Manage. 36(2), 207–227 (Jan 2000)
11. Lucas, W., Topi, H.: Form and function: the impact of query term and operator usage on Web search results. J. Am. Soc. Inf. Sci. Technol. 53(2), 95–108 (Jan 2002)
12. Ogilvie, P., Callan, J.P.: Experiments Using the Lemur Toolkit. In: TREC'01: Proceedings of the 9th Text REtrieval Conference. NIST, Gaithersburg, MD, USA (Feb 2001)
13. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: OSIR'06: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (2006)
14. Palacio, D., Cabanac, G., Sallaberry, C., Hubert, G.: Measuring Effectiveness of Geographic IR Systems in Digital Libraries: Evaluation Framework and Case Study. In: ECDL'10: Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries. LNCS, vol. 6273, pp. 340–351. Springer (Sep 2010)
15. Purday, J.: Think culture: Europeana.eu from concept to construction. The Electronic Library 27(6), 919–937 (2009)
16. Sanderson, M.: Test collection based evaluation of information retrieval systems. Found. Trends Inf. Retr. 4(4), 247–375 (2010)
17. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum 33(1), 6–12 (Sep 1999)
18. Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T.: Searching the web: the public and their queries. J. Am. Soc. Inf. Sci. Technol. 52(3), 226–234 (Feb 2001)
19. Tukey, J.W.: Exploratory data analysis. Addison-Wesley (1977)
20. Voorhees, E.M., Harman, D.K.: Overview of the Seventh Text REtrieval Conference (TREC-7). In: TREC-7: Proceedings of the 7th Text REtrieval Conference. pp. 1–23 (1998)
21. Voorhees, E.M., Harman, D.K.: Overview of the Seventh Text REtrieval Conference (TREC-8). In: TREC-8: Proceedings of the 8th Text REtrieval Conference. pp. 1–23 (1999)
22. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge, MA, USA (2005)
23. White, R.W., Morris, D.: Investigating the querying and browsing behavior of advanced search engine users. In: SIGIR'07: Proceedings of the 30th annual international ACM SIGIR conference. pp. 255–262. ACM, New York, NY, USA (2007)
24. Williamson, D.F., Parker, R.A., Kendrick, J.S.: The box plot: A simple visual method to interpret data. Ann. Intern. Med. 110(11), 916–921 (Jun 1989)