RESQ: Rank-Energy Selective Query Forwarding for Distributed Search Systems

Amin Teymorian
Georgetown University
Washington, DC, USA
amin@cs.georgetown.edu

Xiao Qin
Rutgers University
New Brunswick, NJ, USA
xq26@rutgers.edu

Ophir Frieder
Georgetown University
Washington, DC, USA
ophir@ir.cs.georgetown.edu

ABSTRACT

Selective query forwarding is a promising technique to help scale high-quality and cost-efficient query evaluation in distributed search systems. The basic idea is simple. After a local site receives a query, it determines non-local sites to forward the query to and returns an aggregation of local and non-local results. We introduce "RESQ", a hybrid rank-energy selective query forwarding model. The novel contribution of RESQ is to simultaneously consider both ranking quality and energy costs when making forwarding decisions. Using a large-scale query log and publicly-available energy price time series, we demonstrate the ability of RESQ forwarding to achieve favorable tradeoffs between the possibility of returning high ranking query results and savings in temporally- and spatially-varying energy prices.

Categories and Subject Descriptors

H.3.3 [**Information Storage Systems**]: Information Retrieval System

General Terms

Algorithms, Experimentation, Performance

Keywords

Distributed IR, energy, query forwarding, linear program

1. INTRODUCTION

Scaling high-quality, cost-effective query evaluation is critical to search system performance. In distributed search systems, selective query forwarding helps improve scalability by determining which non-local sites a locally-received query should be forwarded to. A recent approach by Cambazoglu *et al.* [1] increases forwarding efficiency when partial (or "non-replicated") indexes are employed by locally computing an upper bound on the maximum possible ranking of a non-local result for a locally-received query. Based on offline results from non-local indexes, a local site only forwards a query to sites that might have a result that would rank within its top-k results. On the other hand, Kayaaslan *et al.* [6] selectively forward queries in a replicated search system (i.e., evaluations over full indexes) to reduce operating costs based on differences in site energy prices and available query processing capacity. Local sites forward queries to non-local sites with cheaper energy prices with probabilities proportional to the fraction of the local workload that the non-local site could process after evenly distributing its available capacity to sites with higher energy prices.

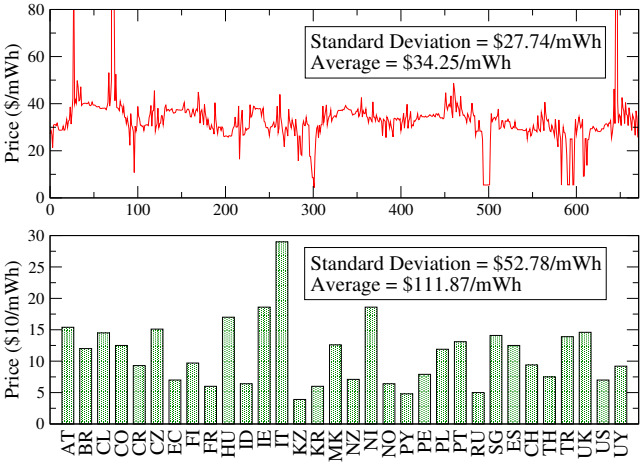

Figure 1: Energy price temporal and spatial variation: (top) NYISO energy prices [7], 15 minute intervals; (bottom) EIA energy prices by country [9].

We propose RESQ (pronounced "rescue"), a hybrid rank-energy selective query forwarding model that obtains the scalability of non-replicated systems and the cost savings of energy-aware approaches. Our idea is to selectively forward queries such that the result ranking guarantee (see Sec. 2) in a non-replicated, geographically-distributed search system is maximized given a budget for spatially- and temporally-varying energy prices, such as those illustrated in Fig. 1. Besides a novel simultaneous consideration of result quality and energy costs, RESQ's design complements existing work well (e.g., dynamic energy prices, cf. [1]; and non-replicated indexes, cf. [6]). Experiments with query and electricity data demonstrate the merit of RESQ's balanced approach (e.g., an 87% ranking guarantee with 46% energy savings).

In the following, we describe RESQ forwarding (Sec. 2), evaluate its performance (Sec. 3), and summarize (Sec. 4).

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

2. RANK-ENERGY SELECTIVE QUERY (RESQ) FORWARDING

We consider a distributed search system consisting of a set of sites $S = \{S_1, S_2, \ldots, S_n\}$ distant geographically and a partitioned index I where each partial index I_i is assigned to a site S_i. When a query q is issued to a local site S_l, $1 \leq l \leq n$, the objective is to select a subset of non-local sites to forward q to such that the sum of the non-local result ranking upper bounds r_i for evaluating q over I_i (or the "ranking guarantee") is maximized given the local site's energy price budget.

2.1 Linear Programming (LP) Formulation

Formulated as a linear program, RESQ forwarding decisions for q maximize the total non-local result rankings:

$$\sum_{i=1}^{n} r_i(q) \times x_i, \tag{1}$$

subject to a per-query, site-specific cost constraint C_l:

$$\sum_{i=1}^{n} c_i(q) \times x_i \leq C_l, \tag{2}$$

where $c_i(q) = \alpha_i \times |q| \times p_i$ (and is explained below), and the feasibility constraints:

$$r_i(q) \geq 0,\ c_i(q) \geq 0,\ \text{and } x_i \in \{0, 1\}. \tag{3}$$

The function $c_i(\cdot)$ is the cost to evaluate a query q with $|q|$ terms over index I_i at site S_i where the spatially- and temporally-varying energy price energy is p_i. A site-specific constant α_i relates units of energy to query complexity. The $x_i \in \{0, 1\}$ indicate whether S_i will be forwarded q (i.e., whether results from S_i will be included in the final aggregate returned results).

For S_l to obtain actual r_i values from non-local site S_j, the query is required to be forwarded and executed (i.e., all sites must execute the query, and the savings from selective query forwarding are obviated). Therefore, selective query forwarding uses $\hat{r}_i$, a locally-computed upper bound on the non-local result rankings (e.g., [1]). However, to simplify notation, we use r_i instead of $\hat{r}_i$. Also, local site S_l necessarily has $x_l = 0$ because all non-local ranking estimates are relative to the results from evaluating q at S_l. A vector of x_i values over all search sites S_i, $1 \leq i \leq n$, represents a solution in RESQ forwarding. The price function may consider non-monetary costs as well (e.g., latency or load capacity). However, as defaulting to local evaluation is possible [6], we limit c_i to electricity-related costs. C is an independent, site-tunable parameter that RESQ only assumes to be in $\mathbb{R}^+$. Consistent with [6], we assume a stable energy price during forwarding and evaluation of an individual query. Lastly, the LP problem is amenable to existing techniques for combinatorial optimization.

2.2 Distributed Algorithm

Upon receiving a query, a (local) search site implements RESQ forwarding in an online, distributed manner. Intuitively, selecting the (non-local) search sites to forward a query can be viewed as a two-step process: locally bounding the non-local result rankings, and optimally selecting the to-be-forwarded-to (non-local) sites based on a tradeoff between ranking quality and energy costs. The RESQ forwarding process is detailed in Algorithm 1, and an example

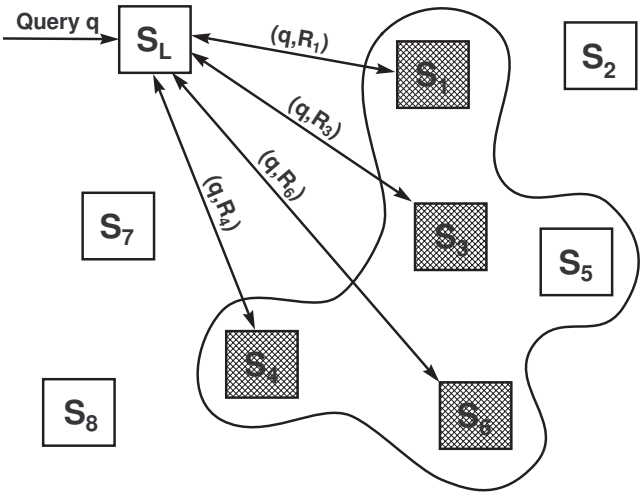

Figure 2: An illustration of RESQ forwarding.

is illustrated in Fig. 2. The algorithm begins when a site (e.g., S_L in Fig. 2) receives a query q at time t. The local site, i.e., the site that received the query, updates the cost information (e.g., p_i and α_i) about all of the non-local sites (e.g., $S_1, \ldots, S_8$ in Fig. 2), and computes the c_is (Line 1). The assumption is that each site receives up-to-date electric-

Algorithm 1 RESQ Forwarding

Require: (q, t), a query q with a timestamp t.
1: $\{c_i(q)\}_{i=1}^{|S|} \leftarrow$ UpdateCosts(t)
2: $\{r_i(q)\}_{i=1}^{|S|} \leftarrow$ BoundRankings(q)
3: $\{x_i\}_{i=1}^{|S|} \leftarrow$ LPSolve$(\{r_i, c_i\}_{i=1}^{|S|}, C)$
4: $R = \emptyset$
5: **for** $i = 1 \rightarrow |S|$ **do**
6: **if** $x_i = 1$ **then**
7: $\hat{R} \leftarrow$ Forward(q, S_i)
8: $R \leftarrow$ Merge$(\hat{R}, R)$
9: **end if**
10: **end for**
11: **return** TopK(R)

ity cost information (e.g., site-to-site broadcast or approximation from historical data [6]). Next, upper bounds on the quality of results that could be returned by non-local sites are locally computed (Line 2). Bounding helps eliminate "false-positive sites," i.e., non-local sites that cannot contribute results that rank higher than those from the local index. For example, in Fig. 2, S_L computes upper bounds (via, for example, [1]) that imply that only S_1, S_3, S_4, S_5, and S_6 (i.e., the sites enclosed in the blob-like boundary) could possibly return at least one result (if they locally evaluated q) with a ranking high enough to make it into the global top-k result set (in contrast to S_2, S_7, and S_8, which have upper bounds on their result rankings that are too low to appear in S_L's local top-k). With the sites guaranteed to be false-positives not selected (i.e., the sites outside the boundary), S_L solves the LP problem as formulated in Sec. 2.1 (Line 3). C denotes the set of constraints from Eq. 2 and Eq. 3. For readability, Algorithm 1 does not reflect a possible new value for $|S|$ after eliminating false-positive sites in the previous line (e.g., the call to BoundRankings reduces

$|S|$ from 8 to 5 implicitly); in our formulation, the simple solution is to append constraints of $x_i = 0$ to C for each excluded site (e.g., in Fig. 2, $x_2 = x_7 = x_8 = 0$). After LPSolve, $x_1 = x_3 = x_4 = x_6 = 1$, which indicates the sites that might contribute results to the global top-k (i.e., the enclosed sites). This subset (i.e., the shaded, enclosed sites) contributes the maximum result ranking upper bounds (i.e., might have at least one document within the top-k ranking) for a budget-restricted energy cost. The global result set R is initialized in Line 4. In Lines 5–10, the query q is forwarded (Line 7) to each member of the subset of plausible non-local sites (Line 6). The results returned to S_L, denoted $\hat{R}$, are then merged into R (Line 8). Finally, in Line 11, the top-k results in R are delivered to the query originator.

3. EVALUATION

The evaluation of RESQ focuses on the tradeoff between ranking quality and energy cost. The reason is that RESQ is a hybrid selective query forwarding algorithm. Using the AOL query log [8] and NYISO electricity prices [7], we make direct comparisons between RESQ, and realistic rank- and energy-only baselines. We describe our experimental setup, methodology, and results in the subsections that follow.

3.1 Setup

We simulate a non-replicated distributed search system that supports the RESQ forwarding model using Java. We use a combination of the AOL query log [8] and New York Independent System Operator (NYISO) day-ahead zonal market electricity pricing data [7]. The log contains approximately $36,000,000$ query events (queries or subsequent result click-throughs) from over $650,000$ users. Although controversial, the AOL query log is publicly-available and used widely (e.g., [2], [3], and [5]), which simplifies comparisons to past and future results. Existing research has also investigated how to effectively anonymize such query logs (e.g., [4]). For privacy, we do not consider the linguistic meaning of the terms, and we do not attempt to deanonymize users. The electricity data contains publicly-available price reports from a particular New York region in 15-minute intervals for a 1-week duration starting on November 7, 2011. Previous studies (e.g., [6]) have used similar data.

For reproducibility, the code, data, and raw figure values from our study are available online at `https://github.com/resqforwarding/resq`.

3.2 Methodology

We train five sites (or data centers) using term frequencies from approximately $18,000,000$ queries, i.e., about half of the query log. The number of data centers is consistent with related studies (e.g., [1]). Testing data are constructed by randomly sampling batches of $1,000,000$ queries from the remaining queries that were not used for training. Each testing set is preprocessed using standard methods. During tests, each data center has a different electricity price series. When a site processes a query, the corresponding electricity price is assigned using the 24-hour time stamp of the query.

Our experiments consist of multiple configurations. Each configuration consists of two settings: a diversity level that represents the variability in the upper bounds on the result rankings from each site, and an energy budget for a local data center that corresponds to C in Eq. 2. We evaluate each configuration using 10 testing sets. The results are averaged, and 95% confidence intervals are computed. The diversity levels produce result rankings that generate 10 average standard deviations over the approximately $1,000,000$ queries in each testing set with a Pearson correlation coefficient of almost 1. The budget values are the first five multiples of the average weekly energy price. Results over all configurations are reported.

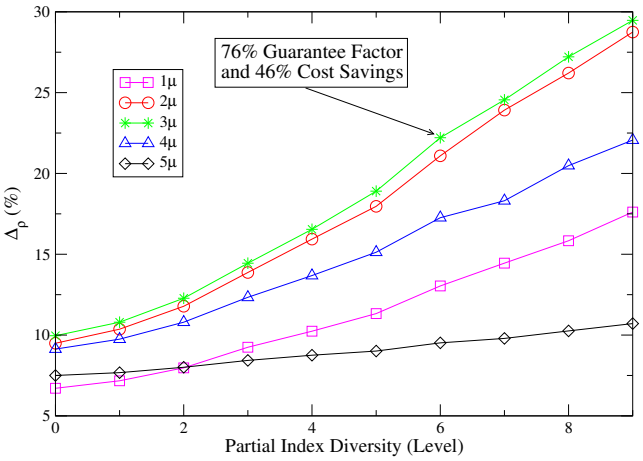

Figure 3: Average performance curves over all configurations for RESQ and baseline B_{rank}.

3.3 Results

We implement RESQ as presented in Sec. 2.2 and evaluate it under each of the configurations. We also evaluate two baseline selective query forwarding algorithms. The first, denoted B_{rank}, represents the ranking-only approach in [1], which is summarized in Sec. 1. The second, denoted B_{energy}, greedily forwards to non-local sites in increasing order of energy price until a cost budget is exceeded.

The performance of RESQ relative to baseline B_{rank} is illustrated in Fig. 3. The horizontal axis indicates the diversity in partial indexes, as described in Sec. 3.2. The vertical axis, denoted Δ_ρ, represents the difference between the ratio of ranking quality upper bounds (or "guarantee factor") and the ratio of cumulative energy cost, denoted ρ_r and ρ_e, respectively. The intuition for combined consideration of both measures is analogous to F-measure. Formally, $\Delta_\rho = \rho_r - \rho_e$, $\rho_r = \frac{r(RESQ)}{r(B_{rank})}$, and $\rho_e = \frac{c(RESQ)}{c(B_{rank})}$. The functions r and c are notational conveniences that refer to the cumulative result rankings and energy costs in Eq. 1 and Eq. 2, respectively, for RESQ or the baselines. A value of $\rho_r = 1$ indicates that result rankings from RESQ are equivalent to B_{rank}. On the other hand, small values of ρ_e indicate low energy costs relative to B_{rank}. Recall that the rank baseline forwards based on ranking quality only. Intuitively, experiments with large positive values of Δ_ρ indicate preservation of ranking quality while reducing energy costs. The results demonstrate that RESQ maintains favorable percentages of the baseline ranking guarantees while reducing energy costs substantially over a range of index diversity levels and energy cost budgets. For example, with a budget of 2μ, i.e., twice the average weekly energy price, RESQ achieves 76% of the original ranking guarantee at 54% of the energy cost. When compared to B_{energy}, energy consumption for a forwarding algorithm should be close to

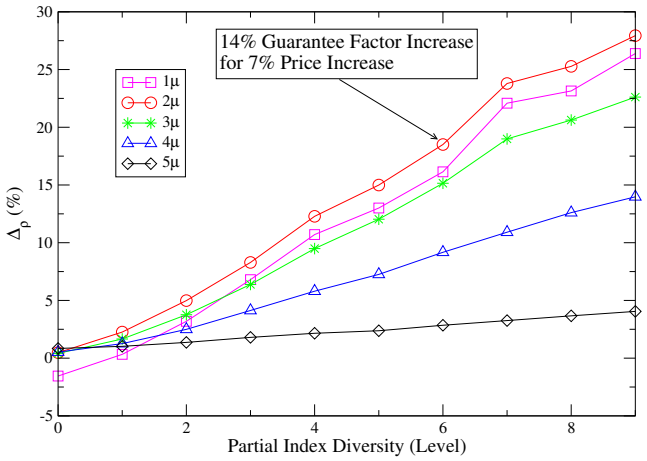

Figure 4: Average performance curves over all configurations for RESQ and baseline B_{energy}.

L \ μ	$B_{rank}(\rho_r)$			$B_{energy}(\rho_r)$		
	5	4	3	5	4	2
0	1.03	0.88	0.69	1.10	1.04	1.00
1	1.04	0.89	0.70	1.09	1.03	0.98
2	1.04	0.90	0.72	1.09	1.03	0.98
3	1.05	0.92	0.74	1.08	0.98	0.92
4	1.05	0.93	0.76	1.08	0.94	0.88
5	1.06	0.95	0.79	1.07	0.92	0.85
6	1.06	0.97	0.82	1.07	0.88	0.81
7	1.07	0.99	0.85	1.06	0.84	0.75
8	1.08	1.01	0.87	1.05	0.82	0.74
9	1.08	1.03	0.90	1.05	0.80	0.71

Table 1: B_{rank} and B_{energy} ρ_r-values for low-, mid-, and high-performing budgets and diversity levels.

the greedy baseline. Under such scenario, the best case is $\rho_e = 1$. The intuition is that the energy cost constraint may be considered to be tuned to a desired cost. On the other hand, an increase in ranking quality over the baseline is still desirable. Therefore, we let $\Delta_\rho = \rho_e - \rho_r$, with $\rho_e = \frac{c(B_{energy})}{c(RESQ)}$ and $\rho_r = \frac{r(B_{energy})}{r(RESQ)}$. As demonstrated by the performance curves in Fig. 4, RESQ leverages relatively small differences in energy costs for substantially increased improvement in ranking quality. For example, under an energy budget of 2μ, RESQ has an energy cost 7% closer to the budget, while increasing the ranking quality by 14%. The improvement is consistent across the parameter settings.

We note that ranking quality thresholding techniques cannot guarantee zero false-positive sites without all non-local sites evaluating a query (which obviates the need for selective query forwarding). Indeed, it is possible that the locally-aggregated results contain only local-results in the global top-k. For example, as mentioned in Sec. 2.2, the non-local sites selected for forwarding might return results that all score lower than documents at the local site. Therefore, the ratio of the result rankings is more relevant to RESQ than the false-positives associated with the underlying thresholding technique (e.g., [1]). We examine RESQ's result ranking ratios in Table 1. The ρ_r values using the baselines B_{rank} and B_{energy} are presented across all diversity levels (denoted L) and energy budgets (denoted μ) under which low-, medium-, and high-performing (as read from left to right, respectively) ranking quality and energy cost tradeoffs were obtained (e.g., ρ_r values for B_{rank} when $\mu = 4$ correspond to the mid-performing curve 4μ in Fig. 3). The steady increase in proportion of the baseline ranking result thresholds is evident across all settings of μ and L. Recall that ρ_r is inverted for B_{energy}, which makes the decreasing values in the right half of the table indicate better performance. Although important to the overall performance illustrated in Fig. 3 and Fig. 4, we omit the ρ_e values because they did not show substantial intra-column variation. The reason is that the coarse granularity of energy budgets relative to site energy prices made sites similarly expensive under a particular budget. In this aspect, the results obtained are pessimistic, as the hybrid nature of RESQ allows for maximal savings under nuanced variance in energy prices or budgets. We also point out that a decrease in ranking guarantee is not necessarily equivalent to a downgrade in ranking quality. Indeed, the results of such systems are a function of retrieved document scores, with properties of the underlying bounds determining the possible quality of results aggregated at a local site.

4. CONCLUSION

We introduced RESQ, a novel hybrid rank-energy selective forwarding model that simultaneously considers ranking quality and energy costs in non-replicated distributed search systems. Experiments with a widely-used query log and publicly-available energy price series demonstrate that RESQ forwarding achieves favorable tradeoffs between the possibility of returning high ranking query results and the costs due to temporally- and spatially-varying energy prices.

5. REFERENCES

[1] B. B. Cambazoglu, E. Varol, E. Kayaaslan, C. Aykanat, and R. Baeza-Yates. Query forwarding in geographically distributed search engines. In *ACM SIGIR '10*.

[2] M. J. Carman, F. Crestani, M. Harvey, and M. Baillie. Towards query log based personalization using topic models. In *ACM CIKM '10*.

[3] S. Gurajada and S. K. P. Index tuning for query-log based on-line index maintenance. In *ACM CIKM '11*.

[4] Y. Hong, X. He, J. Vaidya, N. Adam, and V. Atluri. Effective anonymization of query logs. In *ACM CIKM '09*.

[5] J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *ACM CIKM '09*.

[6] E. Kayaaslan, B. B. Cambazoglu, R. Blanco, F. P. Junqueira, and C. Aykanat. Energy-price-driven query processing in multi-center web search engines. In *ACM SIGIR '11*.

[7] New York Independent System Operator (NYISO). Pricing data. `http://www.nyiso.com/public/markets_operations/market_data/pricing_data/index.jsp`.

[8] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale '06*.

[9] U.S. Energy Information Administration (EIA). Electricity prices, selected countries. `http://www.eia.gov/countries/prices/electricity_industry.cfm`.