A Study on Query Energy Consumption in Web Search Engines

Matteo Catena[1,2] and Nicola Tonellotto[2]

[1] GSSI – Gran Sasso Science Institute, INFN
Viale F.Crispi 7, 67100 L'Aquila, Italy
matteo.catena@gssi.infn.it
[2] National Research Council of Italy,
Via G. Moruzzi 1, 56124 Pisa, Italy
name.surname@isti.cnr.it

Abstract. Commercial web search engines are usually deployed on data centers, which leverage thousands of servers to efficiently answer queries on a large scale. Thanks to these distributed infrastructures, search engines can quickly serve high query volumes. However, the energy consumed by these many servers poses economical and environmental challenges for the Web search engine companies. To tackle such challenges, we advocate the importance of quantifying the energy consumption of a search engine. Therefore, in this study we experimentally analyze energy consumption on a per query basis. Our aim is to evaluate how much energy is consumed by a search server to answer a single query, i.e, its query energy consumption. To perform such measurements, experiments are conducted using the TREC ClueWeb09 collection and the MSN 2006 query log. Results suggest that solving queries require an amount of energy directly proportional to the query processing time.

1 Introduction

Web search engines are today a fundamental part of the Web due to its enormous size. In fact, many users will turn to a search engine to look for the information they need, producing billions of searches every day. Without such a service, the users would have to personally skim through tons of web pages to find what they want. Surfing the Web would be a far less pleasant activity. In spite of the Web hugeness, users are not willing to wait long for queries to be served [1].

In order to satisfy their users, commercial engines continuously crawl and index large amount of web pages, which are promptly retrieved in response to users' queries. To do so, Web search companies – e.g., Google, Yahoo!, Microsoft, Yandex, Baidu, etc. – need computer systems with large computational power and data storage capabilities. Such systems are reported to be composed by thousands and thousands of computers organized in clusters [2], which can efficiently handle big quantities of data. These companies started building large data centers to house such computer clusters. A data center hosts large computer systems together with the associated infrastructures, such as: telecommunications, power supplying, thermal cooling, fire suppression, etc [3].

While data centers enable large-scale search, they raise environmental and economical issues. The ICT sector has been reported to be responsible for roughly 2% of global carbon emissions in 2007, with general purpose data centers accounting for 14% of the ICT footprint [4]. Moreover, power and cooling cost for 15,000 commodity servers could exceed 280,000 $/month in 2003 [2].

For such reasons, improving data centers' energy efficiency has become an attractive and active research area [5]. Nevertheless, little literature exists about energy efficiency in search engines' data centers. Chowdhury is the first to explicitly write about Green Information Retrieval and to propose a research agenda for evaluating and reducing energy consumption in search services [6].

In line with such research agenda, in this work we evaluate the query energy consumption, i.e., the energy consumed by a search server to solve a single query. Such information is important, since energy is costly and commercial Web search engines have to solve each query within a certain monetary budget. Indeed, a search engine should spend between 0.5 cent/query and 2 cent/query to be profitable, including both capital and operational expenditures in this budget [7]. Moreover, recent works try to reduce search engines expense and carbon footprint by taking into direct account their energy consumption. Energy cost has been recently considered for devising energy-saving caching mechanism [8] and to dynamically shift query workloads between multiple, geographically distant, data centers [9,10]. Then, precise measurements of query energy consumptions would be beneficial for such approaches.

2 Experimental Setup & Analysis

Experiments are conducted using the Terrier IR platform [11] on a dedicated Ubuntu 14.04 server; Linux kernel version is 3.13.0-45-generic. The machine is equipped with an 8-core Intel i7-4770K processor and 32GB RAM. To represent the first tier of a Web search engine, we index the ClueWeb09 (Cat. B) document collection. The Porter stemmer is applied to every term and stopwords are removed. Document identifiers and term frequencies are compressed with Elias-Fano encoding [12]. Finally, the inverted index is kept in main memory.

Query terms	1	2	3	4	5	6+
Frequency	316	329	206	79	46	24

Table 1: Query frequencies by number of query terms

We select 1000 unique queries from the MSN 2006 query log. These queries are processed sequentially by a single thread, using BM25 to retrieve the top 1000 documents with WAND [13]. Results are nor stored neither sent over the network, to avoid energy demanding I/O interactions. Moreover, the operating system binds the processing thread on a single core. In this way, only a core is operative while the remaining ones may enter the energy-saving inactive states. Query statistics are reported in Table 1 and Table 2.

For each query, we measure the power consumed by the server. Power consumption is measured at the power socket, using an Alciom PowerSpy2 wattmeter[1]. From each measurement, we remove the power consumed by the server when idle (∼42 Watts), to consider only its dynamic power consumption. Power measurements are taken every 20 milliseconds and are used to compute the query energy consumption. Since some queries take less than 20 milliseconds to be solved, every query is executed 200 times in a row to perform measurements[2]. Query energy consumption is then averaged over the number of executions.

	Min	Q_1	Q_2	Q_3	Max
Processing time	11.46	35.19	114.70	136.50	1368.40

Table 2: Statistics for the query mean processing times (in milliseconds). This table reports the minimum and maximum times observed over the dataset. Moreover, the first (Q_1), second (Q_2) and third (Q_3) quartiles are shown.

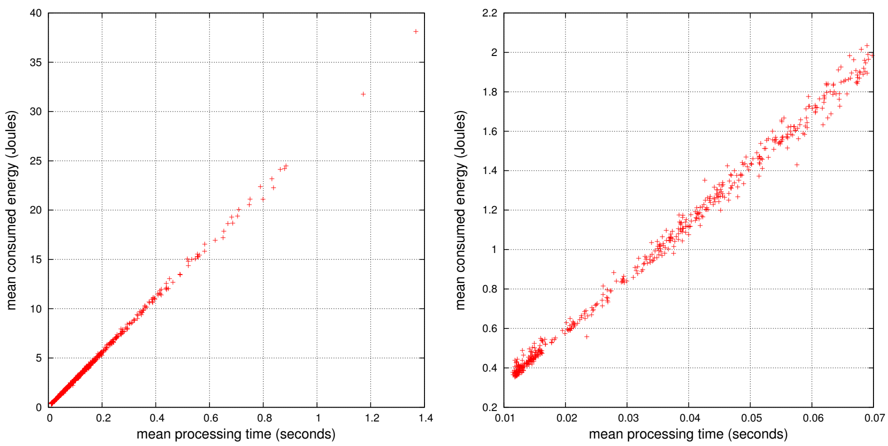

Fig. 1: Each point in the figures represents a query, showing its mean processing time (seconds) on the x-axis and the mean consumed energy (Joules) on the y-axis. The left plot shows the results for every query; the right plot shows just short running ones.

Results are shown in Figure 1, which strongly suggest that the energy consumed for solving a query is linear in its processing time. Indeed, the dynamic power consumed by the server during the experiment is rather stable: the mean absorbed power is 27.55 Watts, with standard deviation 1.81 Watts. The shortest running query requires 0.35 Joules to be solved, while the longest one consumes 38.12 Joules. All in all, the cumulative energy consumption is 3,205.61 Joules.

[1] http://www.alciom.com/en/products/powerspy2-en-gb-2.html
[2] CPU caching effects are not taken into account and left for future work.

3 Conclusions

In this work, we present an experimental setting to measure the energy consumed by a search server to answer a single query, in line with the research agenda proposed in [6]. The experiments results show the query energy consumption to be linear in the query processing time. This indicates that short response times are important for reducing the carbon footprint of search engines. Indeed, low latencies are necessary not only to achieve user satisfaction, but also to tackle the economical and environmental challenges posed by large data centers. Finally, being able to measure the energy consumption of a single query is important, as recent approaches are taking into direct account the search engine power consumption to achieve energy and money savings [8–10]. Therefore, future work should also measure the energy consumed by different query processing stages, i.e., query expansion/reformulation, posting list processing, machine-learned document reordering, snippets generation, etc.

References

1. Schurman, E., Brutlag, J.: Performance related changes and their user impact. In: Proc. Velocity, O'Reilly (2009) 1–1
2. Barroso, L.A., Dean, J., Hölzle, U.: Web Search for a Planet: The Google Cluster Architecture. IEEE Micro **23**(2) (2003) 22–28 IEEE.
3. Hoelzle, U., Barroso, L.A.: The Datacenter As a Computer: An Introduction to the Design of Warehouse-Scale Machines. Morgan&Claypool Publishers (2009)
4. The Climate Group for GeSI: Smart 2020: Enabling the low carbon economy in the information age. `http://goo.gl/4I3vun`, Visited: 2014-08-14
5. Orgerie, A.C., de Assunção, M.D., Lefèvre, L.: A survey on techniques for improving the energy efficiency of large-scale distributed systems. ACM Computing Surveys **46**(4) (2013) 1–35 ACM.
6. Chowdhury, G.: An agenda for green information retrieval research. Information Processing & Management **48**(6) (2012) 1067–1077 Pergamon Press, Inc.
7. Suel, T.: Query optimization in search engines: State of the art and open problems. In: Invited talk at LSDS-IR. (2014)
8. Sazoglu, F.B., Cambazoglu, B.B., Ozcan, R., Altingovde, I.S., Ulusoy, O.: A Financial Cost Metric for Result Caching. In: Proc. SIGIR, ACM (2013) 873–876
9. Kayaaslan, E., Cambazoglu, B.B., Blanco, R., Junqueira, F.P., Aykanat, C.: Energy-price-driven Query Processing in Multi-center Web Search Engines. In: Proc. SIGIR, ACM (2011) 983–992
10. Teymorian, A., Frieder, O., Maloof, M.A.: Rank-energy Selective Query Forwarding for Distributed Search Systems. In: Proc. CIKM, ACM (2013) 389–398
11. Macdonald, C., McCreadie, R., Santos, R.L., Ounis, I.: From puppy to maturity: Experiences in developing terrier. In: Proc. of OSIR at SIGIR, ACM (2012) 60–63
12. Vigna, S.: Quasi-succinct indices. In: Proc. of WSDM, ACM (2013) 83–92
13. Broder, A.Z., Carmel, D., Herscovici, M., Soffer, A., Zien, J.: Efficient query evaluation using a two-level retrieval process. In: Proc. of CIKM, ACM (2003) 426–434