Elastic ChatNoir:
Search Engine for the ClueWeb and the Common Crawl

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast

Bauhaus-Universität Weimar and Leipzig University
<first name>.<last name>@uni-weimar.de and martin.potthast@uni-leipzig.de

Abstract Elastic ChatNoir[1] is an Elasticsearch-based search engine offering a freely accessible search interface for the two ClueWeb corpora and the Common Crawl, together about 3 billion web pages. Running across 130 nodes, Elastic ChatNoir features subsecond response times comparable to commercial search engines. Unlike most commercial search engines, it also offers a powerful API that is available free of charge to IR researchers. Elastic ChatNoir's main purpose is to serve as a baseline for reproducible IR experiments and user studies for the coming years, empowering research at a scale not attainable to many labs beforehand, and to provide a platform for experimenting with new approaches to web search.

1 Introduction

At heart, information retrieval is the art of building the perfect search engine. As a special interest of computer science since 40 years, information retrieval (IR) laid the foundation for technology that society takes for granted today—most notably thanks to the emergence of the world wide web as IR's single most important application domain in terms of deployment scale and commercial success. A large body of work examines nearly every aspect of web search and retrieval under countless practical and theoretical scenarios, contributing as many insights into how (better) web search engines can be built (in the future). For all its rightful claims to fame, there is one thing the scientific community has not yet built for itself: an actual web-scale research search engine.

Wait, what? There are the Googles [1], Lucenes [5], Terriers [7], Indris [10], Galagos [2], and even the non-elastic ChatNoirs [8] of this world, our esteemed readers will interject. True, but none of them have been deployed at web scale, *and* optimized for fast retrieval, *and* made publicly available for free, *and* kept that way. Google was on the right track, but turned commercial and hence opaque; Indri and the non-elastic ChatNoir offer search interfaces to the ClueWeb, but they are not quite as fast as one would have liked them to be, nor capable to withstand a high load of traffic. These shortcomings render the end user search experience unrealistic, since commercial search engines set the bar of users' expectations. Nevertheless, the validity of user studies hinges on realistic user interactions, so that many researchers either index much smaller corpora (e.g., the ClueWeb12-B) at the cost of generalizability, or resort to commercial search engines at the cost of reproducibility and influence over or knowledge about the underlying retrieval model. This applies likewise to experiments where a search engine is automatically queried to reach a higher-level retrieval goal, such as source retrieval for text reuse detection [6]; in fact, commercial search engines hardly offer (affordable) APIs.

[1] Search: www.chatnoir.eu Code: www.github.com/chatnoir-eu

Elastic ChatNoir aims at filling this gap by (1) hosting a freely accessible search engine, (2) indexing the two IR reference corpora ClueWeb09 and ClueWeb12 for compatibility with TREC, (3) indexing at least one instance of the Common Crawl for recency, (4) maintaining at least one baseline retrieval model long-term for reproducibility, (5) offering a free-of-charge API to IR researchers, and (4) publishing its full source code under a permissive open source license.

2 Background and Related Work

The rapid progress of retrieval technology in the 1990s—not least due to the corresponding TREC tracks—were soon adopted by several open-source and research-oriented search projects. Alongside the commercial success of Google [1], modern retrieval models were made available to the wider research community as working software within the popular Lucene library [5], followed by Terrier [7] and Lemur, the latter combining the contributions of the Indri [10] and the Galago [2] teams. In many experiments published at IR conferences and in IR journals, the aforementioned libraries or search systems serve as the underlying retrieval architecture to this day. However, with the rise of ever larger web corpora at TREC, most notably the ClueWeb09 with its more than 500 million English pages, many IR labs lacked the facilities necessary to process and index them. Today, the Common Crawl compiles more than 3 billion pages, with new and potentially even larger versions being published on a monthly basis.

Foreseeing this problem, the Indri team, who crawled the ClueWeb09 and ClueWeb12, also offered on-demand API access to a search engine indexing those corpora. Alas, the search interface has a rather slow retrieval time, rendering it non-realistic, and our demand for API requests swiftly outgrew the quotas. We hence took matters into our own hands and started developing the first version of ChatNoir in 2011 [8], which uses a custom implementation of the BM25F retrieval model [9] as a baseline and indexes the English portion of the ClueWeb09. Sharded across 40 search nodes, we were able to bring down retrieval times to a few seconds for most queries. For five years, the first ChatNoir has supported user studies at scale, shared tasks with dozens of participants, and has served as a teaching subject, answering millions of queries until today. Now the time is ripe for an overhaul and an upgrade.

Twenty years later, our goals do not differ much from Brin's and Page's [1]: "With ~~Google~~ ChatNoir, we have a strong goal to push more development and understanding into the academic realm. Another important design goal was to build systems that reasonable numbers of people can actually use. [..] Our final design goal was to build an architecture that can support novel research activities on large-scale web data, [..] to set up an environment where other researchers can come in [..] and produce interesting results that would have been very difficult to produce otherwise. [..] Another goal we have is to set up a Spacelab-like environment where researchers or even students can propose and do interesting experiments [..]". It goes without saying that today, thanks to technologies that simply did not exist back then, our task is rendered a lot easier.

3 A Scalable Search Engine based on Elasticsearch

With Elastic ChatNoir, we depart from our custom implementation underlying the "old" ChatNoir, adopting Elasticsearch as a well-known, battle-tested, open source search backend, employed by many companies. At the time of writing, Elastic ChatNoir indexes the ClueWeb09, the ClueWeb12, and a 2015 instance of the Common Crawl. Regarding the latter, we plan on updating to the newest version at regular intervals.

Table 1. Key figures of Elastic ChatNoir. Corpora include the ClueWeb09 (cw09), the ClueWeb12 (cw12), and the Common Crawl 11/2015 (cc1511).

Criterion	**Corpus**			$\sum$
	cw09	cw12	cc1511	
Indexed documents	734.8m	638.8m	1.6b	3.0b
Primary shards	40	40	40	120
Shard size	90.0 GB	77.5 GB	242.5 GB	–
Document map files (full)	6.1 TB	6.2 TB	28.8 TB	41.1TB
Index (full)	3.6 TB	3.1 TB	9.7 TB	16.4 TB
Replication	3	3	3	–
Document map files (replicated)	18.3 TB	18.6 TB	86.4 TB	123.3 TB
Index (replicated)	10.7 TB	9.4 TB	29.2 TB	49.3 TB
Total size on disk	29.0 TB	28.0 TB	115.6 TB	172.6 TB

Cluster	**Betaweb**
Nodes	130
Nodes (data)	124
Nodes (master / coord.)	5
Nodes (monitoring)	1
RAM (idle)	5–10 GB
RAM (max)	24 GB
RAM (cache)	150 GB
Avg. query time (warm)	~600 ms
Avg. query time (cold)	~2100 ms

For each corpus, we parse the plain WARC files, heuristically deduce content type and encoding of each entry (since corresponding meta data may be incorrect or missing), assign a deterministic name-based UUID to each entry, and create HDFS map files, mapping UUIDs to a JSON document containing headers and content, and URLs to UUIDs. The map files are input to a Hadoop MapReduce indexing job. In its map phase, from each raw HTML document, the main content is extracted, its language is detected, and meta data such as URLs, keywords, host names, headings, etc. are extracted. Additionally, external meta data (e.g., spam ranks) are mapped to their document IDs. Documents for which no useful content could be extracted, are discarded. During the reduce phase, all information belonging to an individual web page is collected in a multi-field JSON document, which is fed into the Elasticsearch index. As retrieval model for plain text fields, we employ BM25 with various filters and tokenizers for preprocessing. At 124 data nodes and 40 primary shards, an indexing throughput of at least 20,000 documents per second is achieved. For production, each shard is replicated two times.

The web frontend uses the Java Transport API to communicate with Elasticsearch. Every query is first run as a basic filtered Boolean query with AND semantics without expensive operations like proximity, phrasal search, or fuzzy matching. Each shard tries to find up to 70,000 results and then rescores the top-400 results per shard (parameters determined in pilot experiments) with a more complex query, taking into account many more factors (e.g., proximity boosting, additional boosts for Wikipedia articles or home pages, potential penalties for other factors, etc.). Results from the same website are visually grouped and potentially reordered.

Table 1 shows key figures of Elastic ChatNoir and its underlying hardware. Shards are distributed across 8 hard disks per data node. Search efficiency is optimized by continuous “warming,” using the AOL log. This way, Elasticsearch’s node query and request caches are populated, and important parts of the index are served primarily from the host systems’ page cache, guaranteeing fast response times in the order of a few hundred milliseconds for most real-world queries. “Cold” index regions (e.g., non-English queries) are of course slower. The nodes are deployed as Docker containers with different configurations for data and master roles and monitored with Check_MK and Kibana+X-Pack.

Finally, our system comes with a powerful API, accessible with API keys with adjustable quota and access restrictions. Usage of the API and the web interface are logged, allowing for post hoc analyses of user studies on a per-subject basis. API keys are issued for free to interested research parties. All of the above, including code, configuration, and documentation is available open source in our public GitHub repository.

Table 2. Evaluation results for the TREC Web track as ERR@20 / nDCG@20 scores [3,4].

Participant	**TREC Web track (default)**			
(selection)	rank	2013	rank	2014
udel_fang	2	0.176 / 0.310	1	**0.233** / 0.325
ICTNET	4	0.158 / 0.236	2	**0.208** / 0.261
uogTr	3	0.160 / 0.259	5	**0.195** / 0.324
ChatNoir	–	0.130 / 0.193	–	**0.195** / 0.249
Terrier Base	–	–	6	**0.189** / 0.260
udel	5	0.157 / 0.246	7	**0.179** / 0.261
webis / BUW	12	0.101 / 0.181	9	**0.174** / 0.258
wistud	8	0.134 / 0.225	10	**0.174** / 0.291
ut	6	0.152 / 0.228	11	**0.172** / 0.226
Indri Base	14	0.096 / 0.168	12	**0.153** / 0.243

Participant	**TREC Web track (optimized)**			
(selection)	rank	2013	rank	2014
udel_fang	2	0.176 / 0.310	1	**0.233** / 0.325
ChatNoir	–	0.134 / 0.197	–	**0.213** / 0.254
ICTNET	4	0.158 / 0.236	2	**0.208** / 0.261
uogTr	3	0.160 / 0.259	5	**0.195** / 0.324
Terrier Base	–	–	6	**0.189** / 0.260
udel	5	0.157 / 0.246	7	**0.179** / 0.261
webis / BUW	12	0.101 / 0.181	9	**0.174** / 0.258
wistud	8	0.134 / 0.225	10	**0.174** / 0.291
ut	6	0.152 / 0.228	11	**0.172** / 0.226
Indri Base	14	0.096 / 0.168	12	**0.153** / 0.243

4 Evaluation

We evaluated ChatNoir's search effectiveness on the TREC Web tracks of 2013 and 2014 which used the ClueWeb12. One run was conducted using ChatNoir "out of the box," with default parameters, and another run with field weights optimized against previous TREC Web tracks. Table 2 compares our results to a selection of other participants: the default parameters yield a medium rank (though still above the baselines Terrier and Indri), whereas with parameter optimization the performance can be substantially boosted (main optimization criterion was to considerably increase the importance of title matches compared to matches in the text body or other fields).

5 Conclusion

With Elastic ChatNoir, we provide a modern Elasticsearch-based retrieval system for important reference corpora like the ClueWebs and the Common Crawl. ChatNoir is freely available and features a powerful API. In its current version, ChatNoir uses the BM25 retrieval model and achieves subsecond answer times that are similar to commercial search engines, while offering a reasonable retrieval effectiveness as demonstrated by TREC experiments.

In the future, we plan to incorporate further versions of the Common Crawl, so that experiments and user studies with up-to-date web crawls are possible for everyone in a reproducible manner, instead of resorting to commercial search engines as black boxes. Furthermore, we plan to also provide other retrieval models, API functionality of user-defined weighting schemes, and possibly plugin support.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Proc. of WWW 1998.
2. Cartright, M.A., Huston, S., Field, H.: Galago: A modular distributed processing and retrieval system. Proc. of SIGIR 2012 Workshop on Open Source Information Retrieval. pp. 25–31.
3. Collins-Thompson, K., Bennett, P.N., Diaz, F., Clarke, C., Voorhees, E.M.: TREC 2013 web track overview. Proc. of TREC 2013.
4. Collins-Thompson, K., Macdonald, C., Bennett, P.N., Diaz, F., Voorhees, E.M.: TREC 2014 web track overview. Proc. of TREC 2014.
5. Goetz, B.: The Lucene search engine: Powerful, flexible, and free. JavaWorld (2000).
6. Hagen, M., Potthast, M., Adineh, P., Fatehifar, E., Stein, B.: Source retrieval for web-scale text reuse detection. Proc. of CIKM 2017, pp. 2091–2094.
7. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. Proc. of ECIR 2005, pp. 517–519.
8. Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C.: Chatnoir: A search engine for the ClueWeb09 corpus. Proc. of SIGIR 2012. p. 1004.
9. Robertson, S.E., Zaragoza, H., Taylor, M.J.: Simple BM25 extension to multiple weighted fields. Proc. CIKM 2004. pp. 42–49.
10. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. Proc. of ICIA 2005. pp. 2–6.