# ChatNoir: A Search Engine for the ClueWeb09 Corpus

Martin Potthast Matthias Hagen Benno Stein
Jan Graßegger Maximilian Michel Martin Tippmann Clement Welsch
Bauhaus-Universität Weimar
99423 Weimar, Germany
<first name>.<last name>@uni-weimar.de

## ABSTRACT

We present the ChatNoir search engine which indexes the entire English part of the ClueWeb09 corpus. Besides Carnegie Mellon's Indri system, ChatNoir is the second publicly available search engine for this corpus. It implements the classic BM25F information retrieval model including PageRank and spam likelihood. The search engine is scalable and returns the first results within three seconds, which is significantly faster than Indri. A convenient API allows for implementing reproducible experiments based on retrieving documents from the ClueWeb09 corpus. The search engine has successfully accomplished a load test involving 100 000 queries.

**Categories and Subject Descriptors**: H.3.3 [Information Search and Retrieval]: Search process

**General Terms**: Experimentation

**Keywords**: search engine, TREC, ClueWeb09

## 1. INTRODUCTION

Many of the current TREC tracks and TREC style retrieval performance experiments are based on the ClueWeb09 corpus—a collection of 1 billion web pages crawled and provided by the Carnegie Mellon University. As indexing and searching such a large corpus requires a decent amount of hardware probably not available to all researchers interested in TREC style experiments or TREC participation, a public search engine has been provided with the release of the corpus.[1] So far, this is the only publicly available search engine for the ClueWeb09 and it has been used in many TREC runs over the last years. However, the engine is rather slow (answer time of about 10 seconds or more) and only offers the Indri retrieval model (language modeling combined with an inference network). As an alternative, we provide a faster public search engine—ChatNoir.

## 2. CHATNOIR SEARCH

The ChatNoir search engine is based on the classic BM25F retrieval model [4] including the anchor text list provided by the University of Twente [3], the PageRank list provided by the Carnegie Mellon University,[2] and the spam rank list provided by the University of Waterloo [1]. ChatNoir also incorporates an approximate proximity feature with variable width buckets as described by Elsayed et al. [2]. The text body of each document is divided into 64 buckets such that neighboring buckets have a half-bucket overlap. For each keyword, not the exact position is stored in a 1-gram index but occurrence in the individual buckets is indicated via a bit flag. Hence, for each document and each occurring keyword, a 64-bit vector is used in ChatNoir's approximate proximity feature.

The web interface of ChatNoir is similar to that of commercial search engines (snippets, phrasal search, etc.). As for query processing, non-phrasal queries are handled by a 1-gram index of the ClueWeb09 built with Hadoop. Phrasal queries are handled by a 2-gram index and a 3-gram exact position index. For phrase queries containing only 2-grams, the 2-gram index suffices. For phrase queries with longer phrases, the 2-gram index is used to identify documents that contain all 2-grams of a longer phrase while merging the postlists with the 3-gram positional index finally identifies the documents that contain the exact searched phrase. To ensure fast answer times, long queries with more than 2 keywords or phrases are treated in a divide-and-conquer manner. The long query is split into sub-queries for which a parallel retrieval is conducted. The parallel results are then merged into just one list.

The ChatNoir engine runs on a cluster of 10 standard quad-core PCs and 2 eight-core servers. It comes with a web interface and a developer API at `chatnoir.webis.de`. This is the first public alternative to Carnegie Mellon's Indri search for reproducible experiments on the ClueWeb09 without the need of an own cluster for indexing/searching. A load test with 100 000 unique queries from a commercial search engine log showed the robustness and scalability of ChatNoir. The first ten results are typically shown within three seconds compared to more than ten seconds for an Indri search.

## 3. CONCLUSION AND OUTLOOK

With our ChatNoir search engine we provide the second public API for reproducibly searching the ClueWeb09 corpus in TREC style experiments. ChatNoir's much faster BM25F based search offers an alternative to the available Indri retrieval model.

A first use case we are currently handling with ChatNoir is the international PAN competition for automatic plagiarism detection. Therefore, paid human individuals write short reports on TREC topics plagiarizing parts from ClueWeb09 documents found via ChatNoir searches. The competition participants are supposed to find the most likely plagiarism sources also using ChatNoir.

## 4. REFERENCES

[1] Cormack, Smucker, and Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Inf. Retr.* 14(5):441–465, 2011.

[2] Elsayed, Lin, and Metzler. When close enough is good enough: approximate positional indexes for efficient ranked retrieval. *CIKM 2011*, pp. 1993–1996.

[3] Hiemstra and Hauff. MIREX: MapReduce information retrieval experiments. Tech. Report TR-CTIT-10-15, Universiteit Twente, 2010.

[4] Robertson, Zaragoza, and Taylor. Simple BM25 extension to multiple weighted fields. *CIKM 2004*, pp. 42–49.

[1] lemurproject.org/clueweb09.php/index.php#Services

[2] boston.lti.cs.cmu.edu/clueweb09/wiki/tiki-index.php?page=PageRank


*SIGIR'12,* August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.