Supporting Scholarly Search with Keyqueries

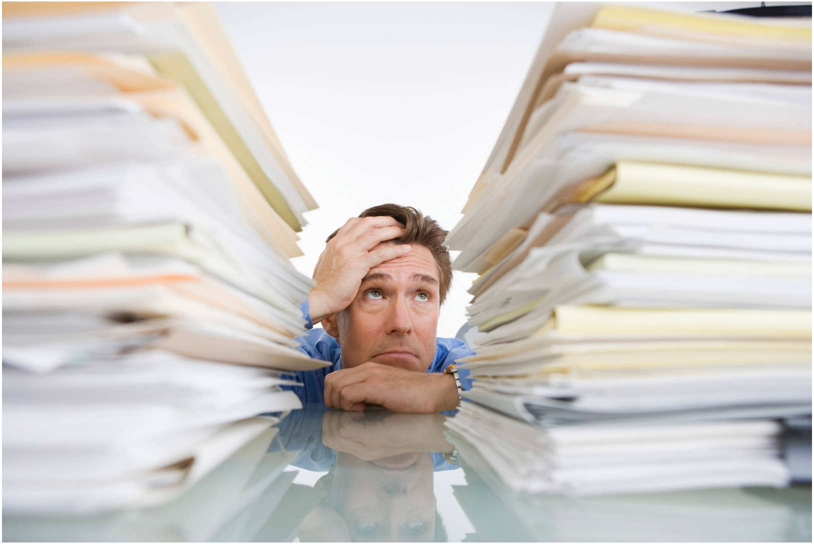
Matthias Hagen Anna Beyer Tim Gollub
Kristof Komlossy Benno Stein

Bauhaus-Universität Weimar
matthias.hagen@uni-weimar.de
@matthias_hagen

ECIR 2016
Padova, Italy
March 23, 2016

WE WANT TO HELP YOU
WITH SCHOLARLY SEARCH

When you start exploring a new topic



When you start exploring a new topic

The first papers are easily found
(colleagues, web search, . . .)

But to find “everything:”

- Follow references and citations
- Check Google Scholar “Related articles”
- Formulate new queries from the read papers

When you start exploring a new topic

The first papers are easily found
(colleagues, web search, . . .)

But to find “everything:”

- Follow references and citations
- Check Google Scholar “Related articles”
- Formulate new queries from the read papers

. . . takes time

When you start exploring a new topic

The first papers are easily found
(colleagues, web search, . . .)

But to find “everything:”

- Follow references and citations
- Check Google Scholar “Related articles”
- Formulate new queries from the read papers

. . . takes time . . . a lot of time

Automatic suggestions for the rescue!

Formalized as a problem

Related Work Search

Given: A small input set D of papers.

Task: Find an output set R of related papers.

Related work for related work search (80+ papers)

Citation-Based	[Golshan et al., SIGMOD 2012]
	[Caragea et al., JCDL 2013]
	[Ekstrand at al., RecSys 2010]
	[Küçüktunç et al., JCDL 2013]
	[Sugiyama and Kan, JCDL 2013]
Content-Based	[Nascimento et al., JCDL 2011]
	[Huang et al., CIKM 2012]
	[Kataria, Mitra, and Bhatia, AAAI 2010]
	[Lu et al., CIKM 2011]
	[Nallapati et al., KDD 2008]
	[Tang et al., PAKDD 2009 & SIGIR 2014]
Mixed	[Google Scholar “Related articles”]
	[El-Arini and Guestrin, KDD 2011]
	[He et al., WWW 2010 & WSDM 2011]
	[Livne et al., SIGIR 2014]
	[Wang and Blei, KDD 2011]

Related work for related work search (80+ papers)

Citation-Based	[Golshan et al., SIGMOD 2012]
	[Caragea et al., JCDL 2013]
	[Ekstrand at al., RecSys 2010]
	[Küçüktunç et al., JCDL 2013]
	[Sugiyama and Kan, JCDL 2013]
Content-Based	[Nascimento et al., JCDL 2011]
	[Huang et al., CIKM 2012]
	[Kataria, Mitra, and Bhatia, AAAI 2010]
	[Lu et al., CIKM 2011]
	[Nallapati et al., KDD 2008]
	[Tang et al., PAKDD 2009 & SIGIR 2014]
Mixed	[Google Scholar "Related articles"]
	[El-Arini and Guestrin, KDD 2011]
	[He et al., WWW 2010 & WSDM 2011]
	[Livne et al., SIGIR 2014]
	[Wang and Blei, KDD 2011]

Our contribution is query formulation (content-based)

The key are . . .

The key are . . . keyqueries

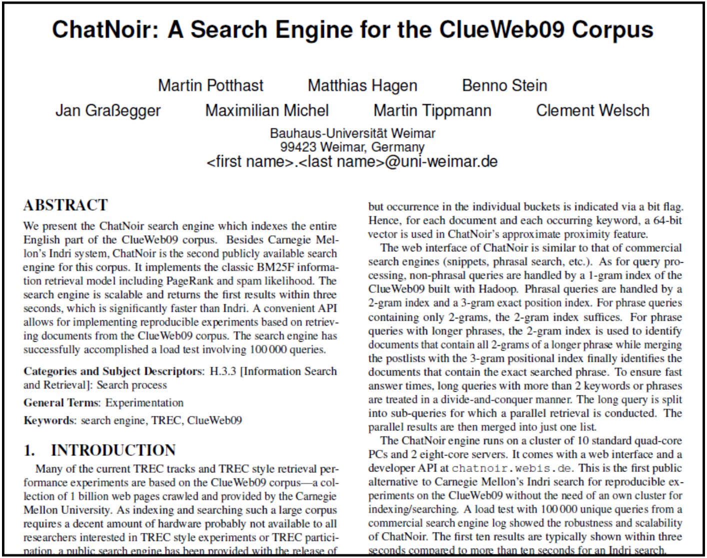
What is a keyquery?

Query q is a keyquery for a set D of documents against a search engine iff

1. Every $d \in D$ is in the top-k results. (specificity)
2. Query q has at least l results. (generality)
3. No $q' \subset q$ satisfies the above. (minimality)

Remark: For small $|D| \leq 5$, typically $l \geq 10$ and $k = 10$.

Example: Keyquery for a paper ($l \geq 1000$, $k = 3$)



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

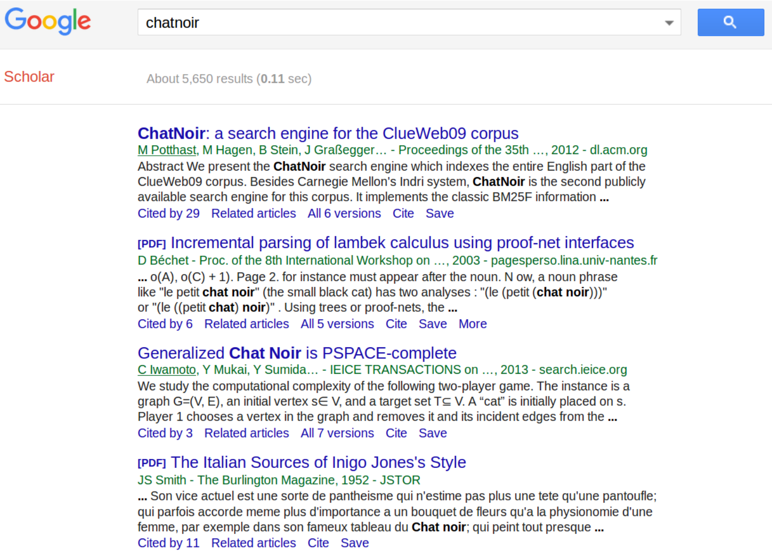
Many of the current TREC tracks and TREC style retrieval performance experiments are based on the ClueWeb09 corpus—a collection of 1 billion web pages crawled and provided by the Carnegie Mellon University. As indexing and searching such a large corpus requires a decent amount of hardware probably not available to all researchers interested in TREC style experiments or TREC participation, a public search engine has been provided with the release of

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.

Example: `chatnoir` is keyquery against Google Scholar



About 5,650 results (**0.11** sec)

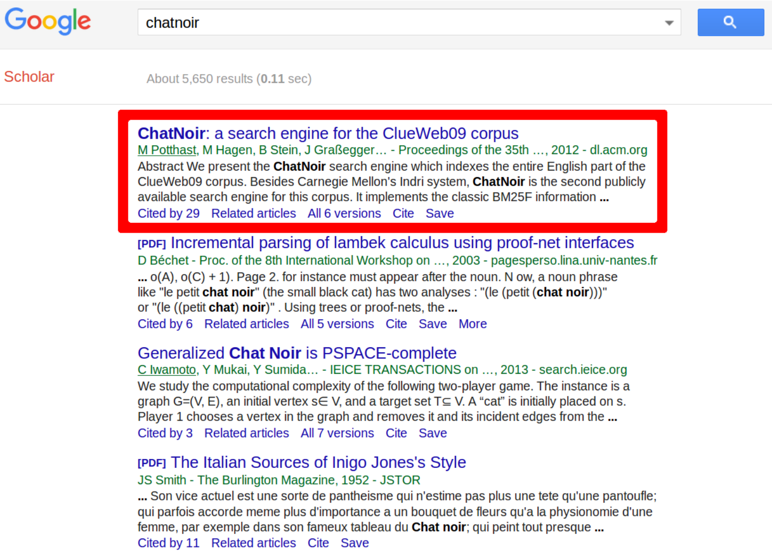
ChatNoir: a search engine for the ClueWeb09 corpus
M Potthast, M Hagen, B Stein, J Graßegger… - Proceedings of the 35th …, 2012 - dl.acm.org
Abstract We present the **ChatNoir** search engine which indexes the entire English part of the ClueWeb09 corpus. Besides Carnegie Mellon's Indri system, **ChatNoir** is the second publicly available search engine for this corpus. It implements the classic BM25F information **...**
Cited by 29 Related articles All 6 versions Cite Save

[PDF] Incremental parsing of lambek calculus using proof-net interfaces
D Béchet - Proc. of the 8th International Workshop on …, 2003 - pagesperso.lina.univ-nantes.fr
... o(A), o(C) + 1). Page 2. for instance must appear after the noun. N ow, a noun phrase like "le petit **chat noir**" (the small black cat) has two analyses : "(le (petit (**chat noir**)))" or "(le ((petit **chat**) **noir**)" . Using trees or proof-nets, the **...**
Cited by 6 Related articles All 5 versions Cite Save More

Generalized **Chat Noir** is PSPACE-complete
C Iwamoto, Y Mukai, Y Sumida… - IEICE TRANSACTIONS on …, 2013 - search.ieice.org
We study the computational complexity of the following two-player game. The instance is a graph G=(V, E), an initial vertex s∈ V, and a target set T⊆ V. A “cat” is initially placed on s. Player 1 chooses a vertex in the graph and removes it and its incident edges from the **...**
Cited by 3 Related articles All 7 versions Cite Save

[PDF] The Italian Sources of Inigo Jones's Style
JS Smith - The Burlington Magazine, 1952 - JSTOR
... Son vice actuel est une sorte de pantheisme qui n'estime pas plus une tete qu'une pantoufle; qui parfois accorde meme plus d'importance a un bouquet de fleurs qu'a la physionomie d'une femme, par exemple dans son fameux tableau du **Chat noir**; qui peint tout presque **...**
Cited by 11 Related articles Cite Save

Example: `chatnoir` is keyquery against Google Scholar



chatnoir

Scholar

About 5,650 results (**0.11** sec)

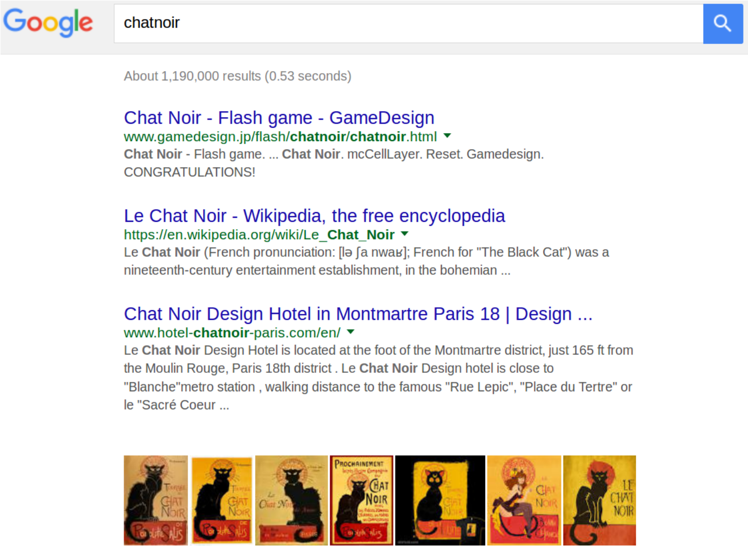
ChatNoir: a search engine for the ClueWeb09 corpus
M Potthast, M Hagen, B Stein, J Graßegger… - Proceedings of the 35th …, 2012 - dl.acm.org
Abstract We present the **ChatNoir** search engine which indexes the entire English part of the ClueWeb09 corpus. Besides Carnegie Mellon's Indri system, **ChatNoir** is the second publicly available search engine for this corpus. It implements the classic BM25F information **...**
Cited by 29 Related articles All 6 versions Cite Save

[PDF] Incremental parsing of lambek calculus using proof-net interfaces
D Béchet - Proc. of the 8th International Workshop on …, 2003 - pagesperso.lina.univ-nantes.fr
... o(A), o(C) + 1). Page 2. for instance must appear after the noun. N ow, a noun phrase like "le petit **chat noir**" (the small black cat) has two analyses : "(le (petit (**chat noir**)))" or "(le ((petit **chat**) **noir**)" . Using trees or proof-nets, the **...**
Cited by 6 Related articles All 5 versions Cite Save More

Generalized **Chat Noir** is PSPACE-complete
C Iwamoto, Y Mukai, Y Sumida… - IEICE TRANSACTIONS on …, 2013 - search.ieice.org
We study the computational complexity of the following two-player game. The instance is a graph G=(V, E), an initial vertex s∈ V, and a target set T⊆ V. A “cat” is initially placed on s. Player 1 chooses a vertex in the graph and removes it and its incident edges from the **...**
Cited by 3 Related articles All 7 versions Cite Save

[PDF] The Italian Sources of Inigo Jones's Style
JS Smith - The Burlington Magazine, 1952 - JSTOR
... Son vice actuel est une sorte de pantheisme qui n'estime pas plus une tete qu'une pantoufle; qui parfois accorde meme plus d'importance a un bouquet de fleurs qu'a la physionomie d'une femme, par exemple dans son fameux tableau du **Chat noir**; qui peint tout presque **...**
Cited by 11 Related articles Cite Save

Example: . . . but not against Google



chatnoir

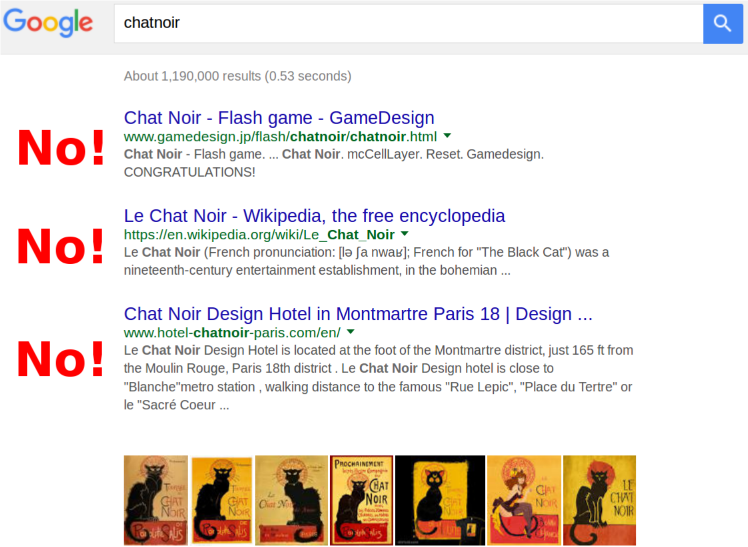
About 1,190,000 results (0.53 seconds)

Chat Noir - Flash game - GameDesign
www.gamedesign.jp/flash/**chatnoir**/**chatnoir**.html
Chat Noir - Flash game. ... **Chat Noir**. mcCellLayer. Reset. Gamedesign. CONGRATULATIONS!

Le Chat Noir - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Le_**Chat**_**Noir**
Le **Chat Noir** (French pronunciation: [lə ʃa nwaʁ]; French for "The Black Cat") was a nineteenth-century entertainment establishment, in the bohemian ...

Chat Noir Design Hotel in Montmartre Paris 18 | Design ...
www.hotel-**chatnoir**-paris.com/en/
Le **Chat Noir** Design Hotel is located at the foot of the Montmartre district, just 165 ft from the Moulin Rouge, Paris 18th district . Le **Chat Noir** Design hotel is close to "Blanche"metro station , walking distance to the famous "Rue Lepic", "Place du Tertre" or le "Sacré Coeur ...

Example: . . . but not against Google



chatnoir

About 1,190,000 results (0.53 seconds)

No!

Chat Noir - Flash game - GameDesign
www.gamedesign.jp/flash/**chatnoir**/**chatnoir**.html ▾
Chat Noir - Flash game. ... **Chat Noir**. mcCellLayer. Reset. Gamedesign. CONGRATULATIONS!

No!

Le Chat Noir - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Le_**Chat**_**Noir** ▾
Le **Chat Noir** (French pronunciation: [lə ʃa nwaʁ]; French for "The Black Cat") was a nineteenth-century entertainment establishment, in the bohemian ...

No!

Chat Noir Design Hotel in Montmartre Paris 18 | Design ...
www.hotel-**chatnoir**-paris.com/en/ ▾
Le **Chat Noir** Design Hotel is located at the foot of the Montmartre district, just 165 ft from the Moulin Rouge, Paris 18th district . Le **Chat Noir** Design hotel is close to "Blanche"metro station , walking distance to the famous "Rue Lepic", "Place du Tertre" or le "Sacré Coeur ...

Keyqueries as a conceptual framework

- Represent a document (set) by its keyqueries
- Related documents also in the top results
- From keywords to keyqueries
- Retrieval model exploited!

Our general algorithmic idea

Assumption: on user side without direct index access, but API

Solution:

1. Keyphrase extraction from input documents [KP-Miner, 2009]
2. Keyquery cover using the keyphrases
3. Keyquery results as suggestions

The keyquery cover problem

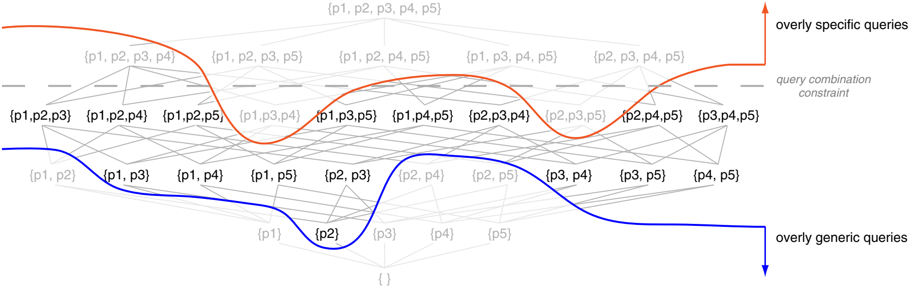
KEYQUERY COVER

Given: (1) A vocabulary W extracted from a set D of documents.
(2) Levels k and l describing keyquery generality.

Task: Find a simple set $Q \subseteq 2^W$ of queries that are keyquery for every $d \in D$ with respect to k and l and that together cover W.

Keyquery cover computation

1. Sort keyphrases by importance
2. Greedily add keyphrases until keyquery
3. Start again with first not-yet-covered phrase



Evaluation

Are the users impressed?!

User study

Collection:	200,000 CS papers (top conferences as seeds)
Search engine:	Lucene 5.0, BM25F (title, abstract, body)
Participants:	13 researchers, 7 students
Topics:	42 provided by participants

1. Participant provides up to five input papers for a familiar topic
2. Participant provides at least one expected document
3. Algorithms run on the input against our collection
4. Participant judges relevance and familiarity

User study results

Algorithm	nDCG@10	rec_{e}@50	rec_{ur}@10
Nascimento	0.58	0.34	0.16
Sofia Search	0.60	0.33	0.20
Google Scholar	0.60	**0.43**	**0.21**
Keyquery Cover	**0.62**	0.37	0.16
KQC+Sofia+Google	**0.65**	**0.48**	**0.24**

- Nascimento query baseline outperformed
- On a par with Google Scholar and Sofia Search
- Rather different suggestions (overlap $<$ 50%)
- Combination most promising

Runtime?!

API requests needed in user study

Nascimento:	19
Google Scholar:	21
Sofia Search:	at least twice as fast as keyqueries
Keyquery Cover:	59

API requests needed in user study

- Nascimento: 19
- Google Scholar: 21
- Sofia Search: at least twice as fast as keyqueries
- Keyquery Cover: 59

Keyqueries could be pre-computed by a scholarly search engine.
Stored in a reverted index. [Pickens, Cooper, and Golovchinsky, CIKM 2010]

Almost the end: The take-home messages!

What we have done

Results

- Keyqueries for scholarly search
- Keyquery cover from keyphrases
- Query baseline outperformed
- On a par with Google Scholar and Sofia Search
- Combination is best

Future Work

- Efficiency
- Other topics and corpora
- Retrieval model influence
- Improved suggestion ranking

What we have (not) done

Results

- Keyqueries for scholarly search
- Keyquery cover from keyphrases
- Query baseline outperformed
- On a par with Google Scholar and Sofia Search
- Combination is best

Future Work

- Efficiency
- Other topics and corpora
- Retrieval model influence
- Improved suggestion ranking

What we have (not) done

Results

- Keyqueries for scholarly search
- Keyquery cover from keyphrases
- Query baseline outperformed
- On a par with Google Scholar and Sofia Search
- Combination is best

Future Work

- Efficiency
- Other topics and corpora
- Retrieval model influence
- Improved suggestion ranking

Thank you

☺