Candidate Document Retrieval for Web-scale Text Reuse Detection

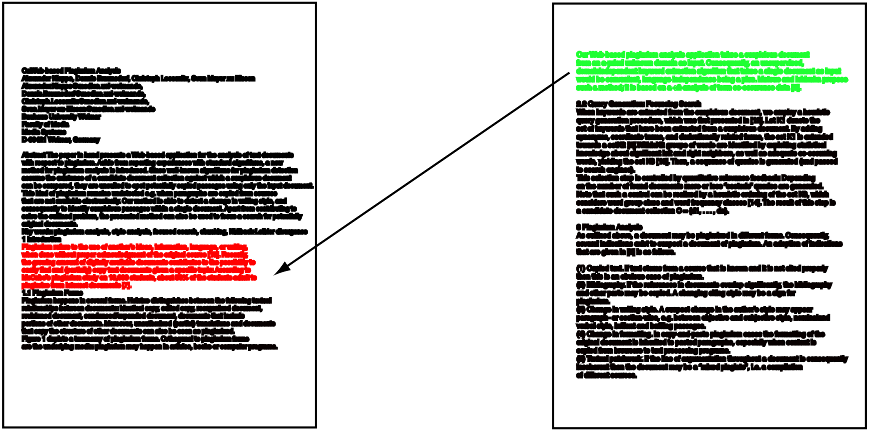
Matthias Hagen   Benno Stein

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

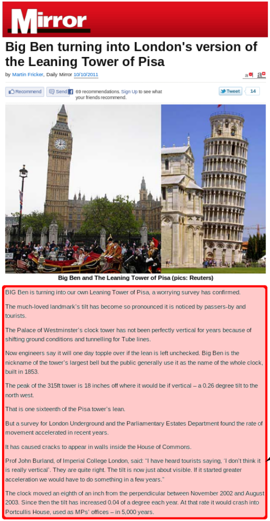
SPIRE 2011
Pisa, Italy
October 19, 2011

Text reuse?

Text from one document used in another.



News reuse



Big Ben turning into London's version of the Leaning Tower of Pisa

by Martin Fricker, Daily Mirror 10/10/2011

69 recommendations. Sign Up to see what your friends recommend.

Big Ben and The Leaning Tower of Pisa (pics: Reuters)

BIG Ben is turning into our own Leaning Tower of Pisa, a worrying survey has confirmed.

The much-loved landmark's tilt has become so pronounced it is noticed by passers-by and tourists.

The Palace of Westminster's clock tower has not been perfectly vertical for years because of shifting ground conditions and tunnelling for Tube lines.

Now engineers say it will one day topple over if the lean is left unchecked. Big Ben is the nickname of the tower's largest bell but the public generally use it as the name of the whole clock, built in 1853.

The peak of the 315ft tower is 18 inches off where it would be if vertical – a 0.26 degree tilt to the north west.

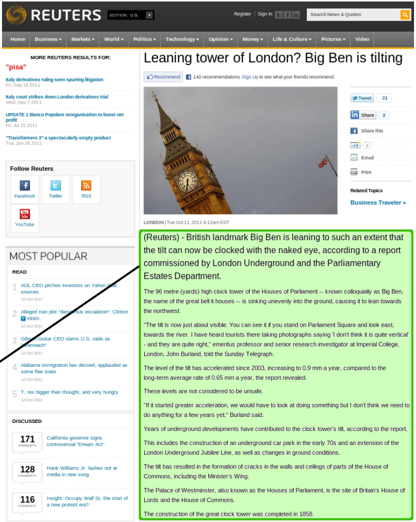
That is one sixteenth of the Pisa tower's lean.

But a survey for London Underground and the Parliamentary Estates Department found the rate of movement accelerated in recent years.

It has caused cracks to appear in walls inside the House of Commons.

Prof John Burland, of Imperial College London, said: "I have heard tourists saying, 'I don't think it is really vertical'. They are quite right. The tilt is now just about visible. If it started greater acceleration we would have to do something in a few years."

The clock moved an eighth of an inch from the perpendicular between November 2002 and August 2003. Since then the tilt has increased 0.04 of a degree each year. At that rate it would crash into Portcullis House, used as MPs' offices – in 5,000 years.



MORE REUTERS RESULTS FOR:

"pisa"

Italy derivatives ruling seen spurring litigation
Fri, Sep 16 2011

Italy court strikes down London derivatives trial
Wed, Sep 7 2011

UPDATE 1-Banco Popolare reorganisation to boost net profit
Fri, Jul 15 2011

"Transformers 3" a spectacularly empty product
Tue, Jun 28 2011

Follow Reuters

MOST POPULAR

READ

1 AOL CEO pitches investors on Yahoo deal: sources
12 Oct 2011

2 Alleged Iran plot "dangerous escalation": Clinton
VIDEO
12 Oct 2011

3 Gibson Guitar CEO slams U.S. raids as "overreach"
12 Oct 2011

4 Alabama immigration law decried, applauded as some flee state
12 Oct 2011

5 T. rex bigger than thought, and very hungry
12 Oct 2011

DISCUSSED

171 COMMENTS	California governor signs controversial "Dream Act"
128 COMMENTS	Hank Williams Jr. lashes out at media in new song
116 COMMENTS	Insight: Occupy Wall St, the start of a new protest era?

Leaning tower of London? Big Ben is tilting

140 recommendations. Sign Up to see what your friends recommend.

Related Topics

Business Traveler »

LONDON | Tue Oct 11, 2011 6:12am EDT

(Reuters) - British landmark Big Ben is leaning to such an extent that the tilt can now be clocked with the naked eye, according to a report commissioned by London Underground and the Parliamentary Estates Department.

The 96 metre (yards) high clock tower of the Houses of Parliament -- known colloquially as Big Ben, the name of the great bell it houses -- is sinking unevenly into the ground, causing it to lean towards the northwest.

"The tilt is now just about visible. You can see it if you stand on Parliament Square and look east, towards the river. I have heard tourists there taking photographs saying 'I don't think it is quite vertical' - and they are quite right," emeritus professor and senior research investigator at Imperial College, London, John Burland, told the Sunday Telegraph.

The level of the tilt has accelerated since 2003, increasing to 0.9 mm a year, compared to the long-term average rate of 0.65 mm a year, the report revealed.

These levels are not considered to be unsafe.

"If it started greater acceleration, we would have to look at doing something but I don't think we need to do anything for a few years yet," Burland said.

Years of underground developments have contributed to the clock tower's tilt, according to the report.

This includes the construction of an underground car park in the early 70s and an extension of the London Underground Jubilee Line, as well as changes in ground conditions.

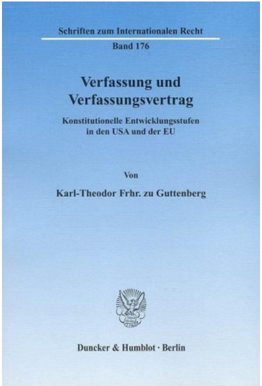
The tilt has resulted in the formation of cracks in the walls and ceilings of parts of the House of Commons, including the Minister's Wing.

The Palace of Westminster, also known as the Houses of Parliament, is the site of Britain's House of Lords and the House of Commons.

The construction of the great clock tower was completed in 1858.

Plagiarism

Karl-Theodor zu Guttenberg
(former German Minister of Defence)



Schriften zum Internationalen Recht
Band 176

Verfassung und
Verfassungsvertrag

Konstitutionelle Entwicklungsstufen
in den USA und der EU

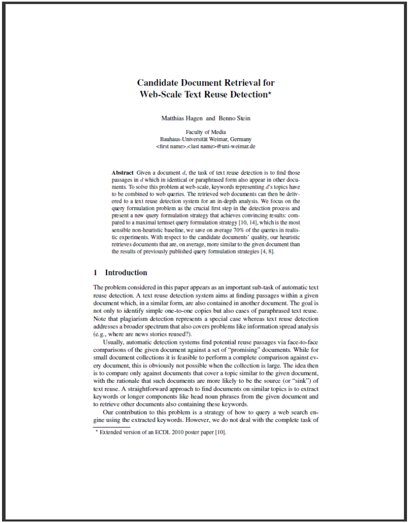
Von

Karl-Theodor Frhr. zu Guttenberg

Duncker & Humblot · Berlin

60% of dissertation plagiarized

Paper versions



Candidate Document Retrieval for Web-Scale Text Reuse Detection*

Matthias Hagen and Benno Stein

Faculty of Media
Bauhaus-Universität Weimar, Germany
<first name>.<last name>@uni-weimar.de

Abstract Given a document d, the task of text reuse detection is to find those passages in d which in identical or paraphrased form also appear in other documents. To solve this problem at web-scale, keywords representing d's topics have to be combined to web queries. The retrieved web documents can then be delivered to a text reuse detection system for an in-depth analysis. We focus on the query formulation problem as the crucial first step in the detection process and present a new query formulation strategy that achieves convincing results: compared to a maximal termset query formulation strategy [10, 14], which is the most sensible non-heuristic baseline, we save on average 70% of the queries in realistic experiments. With respect to the candidate documents' quality, our heuristic retrieves documents that are, on average, more similar to the given document than the results of previously published query formulation strategies [4, 8].

1 Introduction

The problem considered in this paper appears as an important sub-task of automatic text reuse detection. A text reuse detection system aims at finding passages within a given document which, in a similar form, are also contained in another document. The goal is not only to identify simple one-to-one copies but also cases of paraphrased text reuse. Note that plagiarism detection represents a special case whereas text reuse detection addresses a broader spectrum that also covers problems like information spread analysis (e.g., where are news stories reused?).

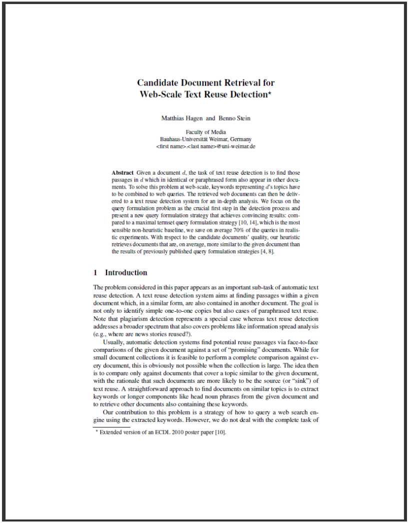
Usually, automatic detection systems find potential reuse passages via face-to-face comparisons of the given document against a set of "promising" documents. While for small document collections it is feasible to perform a complete comparison against every document, this is obviously not possible when the collection is large. The idea then is to compare only against documents that cover a topic similar to the given document, with the rationale that such documents are more likely to be the source (or "sink") of text reuse. A straightforward approach to find documents on similar topics is to extract keywords or longer components like head noun phrases from the given document and to retrieve other documents also containing these keywords.

Our contribution to this problem is a strategy of how to query a web search engine using the extracted keywords. However, we do not deal with the complete task of

* Extended version of an ECDL 2010 poster paper [10].

SPIRE 2011 full paper

Paper versions



Candidate Document Retrieval for Web-Scale Text Reuse Detection*

Matthias Hagen and Benno Stein

Faculty of Media
Bauhaus-Universität Weimar, Germany
<first name>.<last name>@uni-weimar.de

Abstract Given a document d, the task of text reuse detection is to find those passages in d which in identical or paraphrased form also appear in other documents. To solve this problem at web-scale, keywords representing d's topics have to be combined to web queries. The retrieved web documents can then be delivered to a text reuse detection system for an in-depth analysis. We focus on the query formulation problem as the crucial first step in the detection process and present a new query formulation strategy that achieves convincing results: compared to a maximal termset query formulation strategy [10, 14], which is the most sensible non-heuristic baseline, we save on average 70% of the queries in realistic experiments. With respect to the candidate documents' quality, our heuristic retrieves documents that are, on average, more similar to the given document than the results of previously published query formulation strategies [4, 8].

1 Introduction

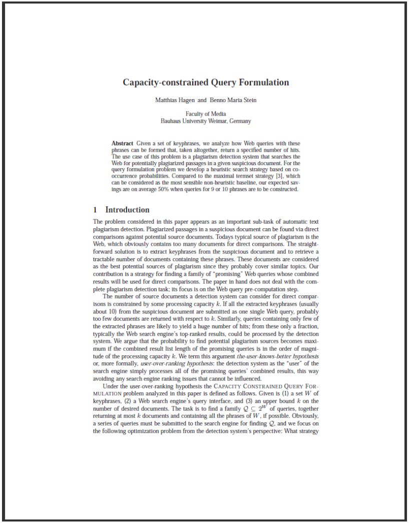
The problem considered in this paper appears as an important sub-task of automatic text reuse detection. A text reuse detection system aims at finding passages within a given document which, in a similar form, are also contained in another document. The goal is not only to identify simple one-to-one copies but also cases of paraphrased text reuse. Note that plagiarism detection represents a special case whereas text reuse detection addresses a broader spectrum that also covers problems like information spread analysis (e.g., where are news stories reused?).

Usually, automatic detection systems find potential reuse passages via face-to-face comparisons of the given document against a set of "promising" documents. While for small document collections it is feasible to perform a complete comparison against every document, this is obviously not possible when the collection is large. The idea then is to compare only against documents that cover a topic similar to the given document, with the rationale that such documents are more likely to be the source (or "sink") of text reuse. A straightforward approach to find documents on similar topics is to extract keywords or longer components like head noun phrases from the given document and to retrieve other documents also containing these keywords.

Our contribution to this problem is a strategy of how to query a web search engine using the extracted keywords. However, we do not deal with the complete task of

* Extended version of an ECDL 2010 poster paper [10].

SPIRE 2011 full paper



Capacity-constrained Query Formulation

Matthias Hagen and Benno Maria Stein

Faculty of Media
Bauhaus University Weimar, Germany

Abstract Given a set of keyphrases, we analyze how Web queries with these phrases can be formed that, taken altogether, return a specified number of hits. The use case of this problem is a plagiarism detection system that searches the Web for potentially plagiarized passages in a given suspicious document. For the query formulation problem we develop a heuristic search strategy based on co-occurrence probabilities. Compared to the maximal termset strategy [3], which can be considered as the most sensible non-heuristic baseline, our expected savings are on average 50% when queries for 9 or 10 phrases are to be constructed.

1 Introduction

The problem considered in this paper appears as an important sub-task of automatic text plagiarism detection. Plagiarized passages in a suspicious document can be found via direct comparisons against potential source documents. Todays typical source of plagiarism is the Web, which obviously contains too many documents for direct comparisons. The straightforward solution is to extract keyphrases from the suspicious document and to retrieve a tractable number of documents containing these phrases. These documents are considered as the best potential sources of plagiarism since they probably cover similar topics. Our contribution is a strategy for finding a family of "promising" Web queries whose combined results will be used for direct comparisons. The paper in hand does not deal with the complete plagiarism detection task; its focus is on the Web query pre-computation step.

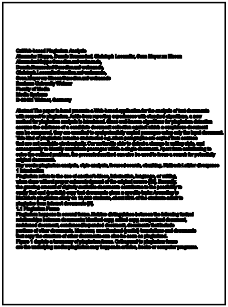
The number of source documents a detection system can consider for direct comparisons is constrained by some processing capacity k. If all the extracted keyphrases (usually about 10) from the suspicious document are submitted as one single Web query, probably too few documents are returned with respect to k. Similarly, queries containing only few of the extracted phrases are likely to yield a huge number of hits; from these only a fraction, typically the Web search engine's top-ranked results, could be processed by the detection system. We argue that the probability to find potential plagiarism sources becomes maximum if the combined result list length of the promising queries is in the order of magnitude of the processing capacity k. We term this argument *the-user-knows-better hypothesis* or, more formally, *user-over-ranking hypothesis*: the detection system as the "user" of the search engine simply processes all of the promising queries' combined results, this way avoiding any search engine ranking issues that cannot be influenced.

Under the user-over-ranking hypothesis the CAPACITY CONSTRAINED QUERY FORMULATION problem analyzed in this paper is defined as follows. Given is (1) a set W of keyphrases, (2) a Web search engine's query interface, and (3) an upper bound k on the number of desired documents. The task is to find a family $\mathcal{Q} \subseteq 2^W$ of queries, together returning at most k documents and containing all the phrases of W, if possible. Obviously, a series of queries must be submitted to the search engine for finding $\mathcal{Q}$, and we focus on the following optimization problem from the detection system's perspective: What strategy

ECDL 2010 poster

Text reuse detection

Given "suspicious" document



On Web-based Plagiarism Analysis
Alexander Kleppe, Dennis Braunsdorf, Christoph Loessnitz, Sven Meyer zu Eissen
Alexander.Kleppe@medien.uni-weimar.de,
Dennis.Braunsdorf@medien.uni-weimar.de,
Christoph.Loessnitz@medien.uni-weimar.de,
Sven.Meyer-zu-Eissen@medien.uni-weimar.de
Bauhaus University Weimar
Faculty of Media
Media Systems
D-99421 Weimar, Germany

Abstract The paper in hand presents a Web-based application for the analysis of text documents with respect to plagiarism. Aside from reporting experiences with standard algorithms, a new method for plagiarism analysis is introduced. Since well-known algorithms for plagiarism detection assume the existence of a candidate document collection against which a suspicious document can be compared, they are unsuited to spot potentially copied passages using only the input document. This kind of plagiarism remains undetected e.g. when paragraphs are copied from sources that are not available electronically. Our method is able to detect a change in writing style, and consequently to identify suspicious passages within a single document. Apart from contributing to solve the outlined problem, the presented method can also be used to focus a search for potentially original documents.

Key words: plagiarism analysis, style analysis, focused search, chunking, Kullback-Leibler divergence

1 Introduction

Plagiarism refers to the use of another's ideas, information, language, or writing, when done without proper acknowledgment of the original source [15]. Recently, the growing amount of digitally available documents contributes to the possibility to easily find and (partially) copy text documents given a specific topic. According to McCabe's plagiarism study on 18,000 students, about 50% of the students admit to plagiarize from Internet documents [7].

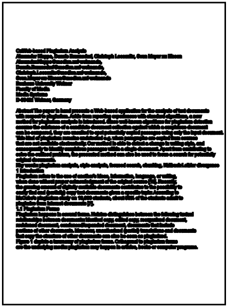
1.1 Plagiarism Forms

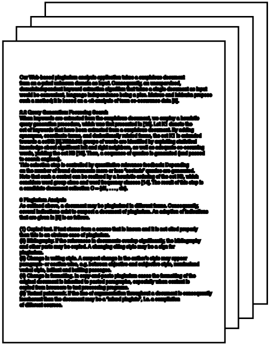
Plagiarism happens in several forms. Heintze distinguishes between the following textual relationships between documents: identical copy, edited copy, reorganized document, revisioned document, condensed/expanded document, documents that include portions of other documents. Moreover, unauthorized (partial) translations and documents that copy the structure of other documents can also be seen as plagiarized. Figure 1 depicts a taxonomy of plagiarism forms. Orthogonal to plagiarism forms are the underlying media: plagiarism may happen in articles, books or computer programs.

Text reuse detection

Given "suspicious" document

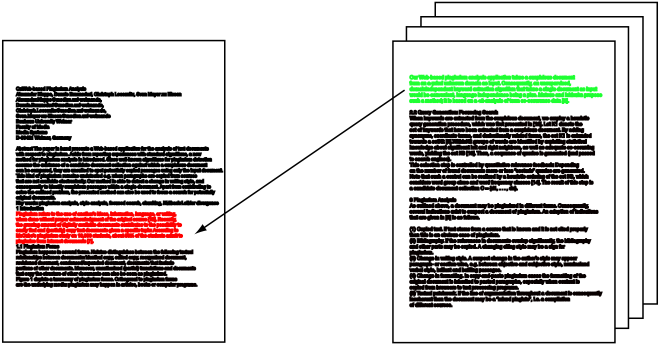
Step 1: Find a set of candidate documents





Text reuse detection

Given	"suspicious" document
Step 1:	Find a set of candidate documents
Step 2:	In-depth analysis against each candidate



We focus on Step 1

Candidate document retrieval

Candidate document retrieval

Observations

- Text reuse source = the entire Web → web search
- Same topic doc's = more likely source

Candidate document retrieval

Observations

- Text reuse source = the entire Web
- Same topic doc's = more likely source

→ web search

- Too many candidates = bad runtime
- Up to k candidates = reasonable runtime

→ system capacity k

Candidate document retrieval

Observations

Text reuse source	= the entire Web	→ web search
Same topic doc's	= more likely source	
Too many candidates	= bad runtime	→ system capacity k
Up to k candidates	= reasonable runtime	

Idea

Retrieve a feasible number of similar web documents.

Standing on the shoulders of . . .

Random string as query	[Dasdan et al., CIKM 2009]
Rare keywords as query	[Dasdan et al., CIKM 2009]
Important keywords as query	[Yang et al., WSDM 2009] [Bendersky and Croft, WSDM 2009]

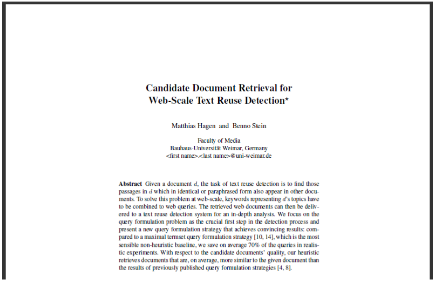
What query to formulate from important keywords?

Example

information retrieval   text reuse

detection system   web search   query formulation

capacity constrained   search engine



Candidate Document Retrieval for Web-Scale Text Reuse Detection*

Matthias Hagen and Benno Stein

Faculty of Media
Bauhaus-Universität Weimar, Germany
<first name>.<last name>@uni-weimar.de

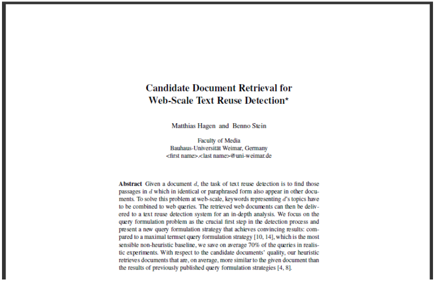
Abstract Given a document d, the task of text reuse detection is to find those passages in d which in identical or paraphrased form also appear in other documents. To solve this problem at web-scale, keywords representing d's topics have to be combined to web queries. The retrieved web documents can then be delivered to a text reuse detection system for an in-depth analysis. We focus on the query formulation problem as the crucial first step in the detection process and present a new query formulation strategy that achieves convincing results: compared to a maximal termset query formulation strategy [10, 14], which is the most sensible non-heuristic baseline, we save on average 70% of the queries in realistic experiments. With respect to the candidate documents' quality, our heuristic retrieves documents that are, on average, more similar to the given document than the results of previously published query formulation strategies [4, 8].

Example

information retrieval   text reuse

detection system   web search   query formulation

capacity constrained   search engine

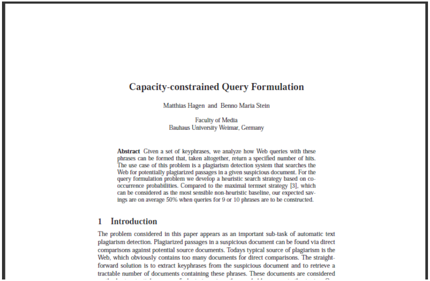


Candidate Document Retrieval for Web-Scale Text Reuse Detection*

Matthias Hagen and Benno Stein

Faculty of Media
Bauhaus-Universität Weimar, Germany
<first name>.<last name>@uni-weimar.de

Abstract Given a document *d*, the task of text reuse detection is to find those passages in *d* which in identical or paraphrased form also appear in other documents. To solve this problem at web-scale, keywords representing *d*'s topics have to be combined to web queries. The retrieved web documents can then be delivered to a text reuse detection system for an in-depth analysis. We focus on the query formulation problem as the crucial first step in the detection process and present a new query formulation strategy that achieves convincing results: compared to a maximal termset query formulation strategy [10, 14], which is the most sensible non-heuristic baseline, we save on average 70% of the queries in realistic experiments. With respect to the candidate documents' quality, our heuristic retrieves documents that are, on average, more similar to the given document than the results of previously published query formulation strategies [4, 8].



Capacity-constrained Query Formulation

Matthias Hagen and Benno Maria Stein

Faculty of Media
Bauhaus University Weimar, Germany

Abstract Given a set of keyphrases, we analyze how Web queries with these phrases can be formed that, taken altogether, return a specified number of hits. The use case of this problem is a plagiarism detection system that searches the Web for potentially plagiarized passages in a given suspicious document. For the query formulation problem we develop a heuristic search strategy based on co-occurrence probabilities. Compared to the maximal termset strategy [3], which can be considered as the most sensible non-heuristic baseline, our expected savings are on average 50% when queries for 9 or 10 phrases are to be constructed.

1 Introduction

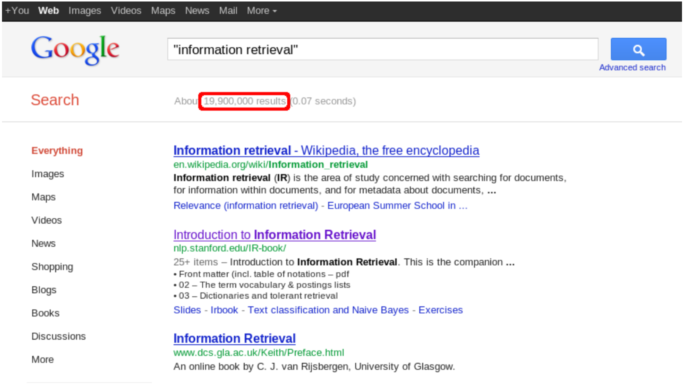
The problem considered in this paper appears as an important sub-task of automatic text plagiarism detection. Plagiarized passages in a suspicious document can be found via direct comparisons against potential source documents. Todays typical source of plagiarism is the Web, which obviously contains too many documents for direct comparisons. The straightforward solution is to extract keyphrases from the suspicious document and to retrieve a tractable number of documents containing these phrases. These documents are considered

Single keyword queries?

information retrieval ~~text reuse~~

~~detection system~~ ~~web search~~ ~~query formulation~~

~~capacity constrained~~ ~~search engine~~



"information retrieval"

About 19,900,000 results (0.07 seconds)

Information retrieval - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Information_retrieval**
Information retrieval (**IR**) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, **...**
Relevance (information retrieval) - European Summer School in ...

Introduction to **Information Retrieval**
nlp.stanford.edu/IR-book/
25+ items – Introduction to **Information Retrieval**. This is the companion **...**

- Front matter (incl. table of notations – pdf
- 02 – The term vocabulary & postings lists
- 03 – Dictionaries and tolerant retrieval

Slides - Irbook - Text classification and Naive Bayes - Exercises

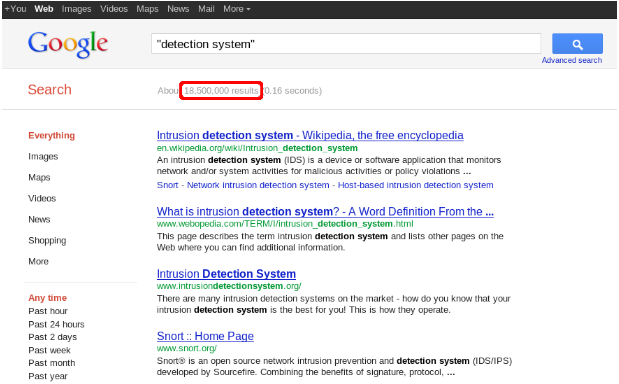
Information Retrieval
www.dcs.gla.ac.uk/Keith/Preface.html
An online book by C. J. van Rijsbergen, University of Glasgow.

Single keyword queries?

~~information retrieval~~ ~~text reuse~~

detection system ~~web search~~ ~~query formulation~~

~~capacity constrained~~ ~~search engine~~



"detection system"

About 18,500,000 results (0.16 seconds)

Intrusion **detection system** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Intrusion_**detection**_**system**
An intrusion **detection system** (IDS) is a device or software application that monitors network and/or system activities for malicious activities or policy violations **...**
Snort - Network intrusion detection system - Host-based intrusion detection system

What is intrusion **detection system**? - A Word Definition From the **...**
www.webopedia.com/TERM/I/intrusion_**detection**_**system**.html
This page describes the term intrusion **detection system** and lists other pages on the Web where you can find additional information.

Intrusion **Detection System**
www.intrusion**detectionsystem**.org/
There are many intrusion detection systems on the market - how do you know that your intrusion **detection system** is the best for you! This is how they operate.

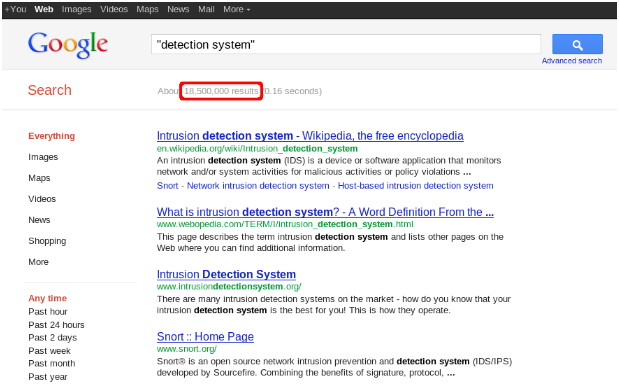
Snort :: Home Page
www.snort.org/
Snort® is an open source network intrusion prevention and **detection system** (IDS/IPS) developed by Sourcefire. Combining the benefits of signature, protocol, **...**

Single keyword queries? Underspecific!

~~information retrieval~~ ~~text reuse~~

detection system ~~web search~~ ~~query formulation~~

~~capacity constrained~~ ~~search engine~~



"detection system"

About 18,500,000 results (0.16 seconds)

Intrusion **detection system** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Intrusion_**detection**_**system**
An intrusion **detection system** (IDS) is a device or software application that monitors network and/or system activities for malicious activities or policy violations **...**
Snort - Network intrusion detection system - Host-based intrusion detection system

What is intrusion **detection system**? - A Word Definition From the **...**
www.webopedia.com/TERM/I/intrusion_**detection**_**system**.html
This page describes the term intrusion **detection system** and lists other pages on the Web where you can find additional information.

Intrusion **Detection System**
www.intrusion**detectionsystem**.org/
There are many intrusion detection systems on the market - how do you know that your intrusion **detection system** is the best for you! This is how they operate.

Snort :: Home Page
www.snort.org/
Snort® is an open source network intrusion prevention and **detection system** (IDS/IPS) developed by Sourcefire. Combining the benefits of signature, protocol, **...**

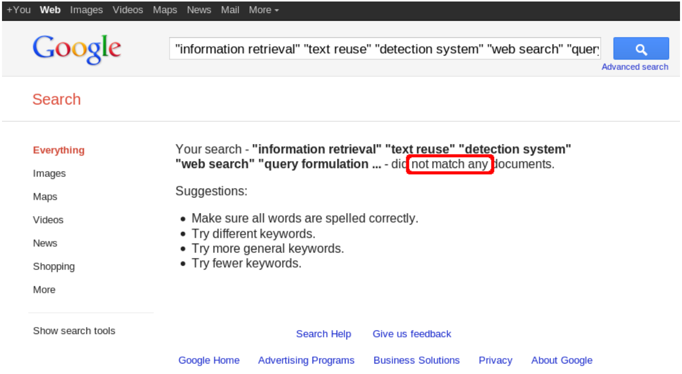
All keywords at once?

[Yang et al., WSDM 2009]

information retrieval text reuse

detection system web search query formulation

capacity constrained search engine



Your search - **"information retrieval" "text reuse" "detection system" "web search" "query formulation ...** - did not match any documents.

Suggestions:

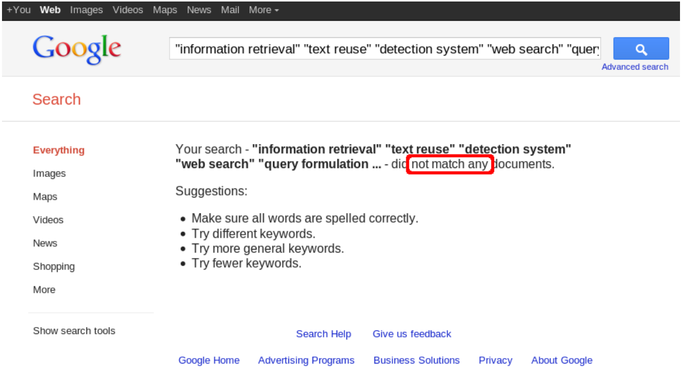
- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

All keywords at once? Overspecific!

`information retrieval` `text reuse`

`detection system` `web search` `query formulation`

`capacity constrained` `search engine`



"information retrieval" "text reuse" "detection system" "web search" "quer

Your search - **"information retrieval" "text reuse" "detection system" "web search" "query formulation ...** - did not match any documents.

Suggestions:

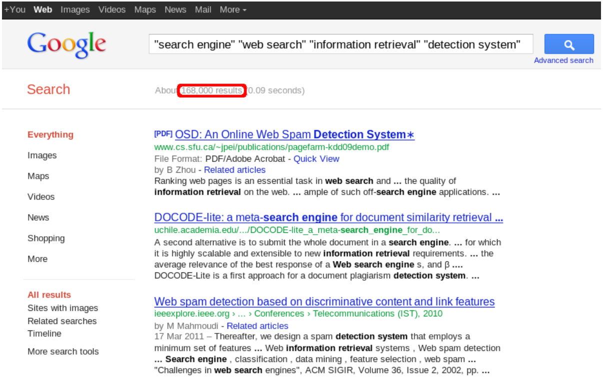
- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Delete rare keywords till *k* results? [Bendersky and Croft, WSDM 2009]

information retrieval ~~text reuse~~

detection system web search ~~query formulation~~

~~capacity constrained~~ search engine



"search engine" "web search" "information retrieval" "detection system"

About 168,000 results (0.09 seconds)

[PDF] OSD: An Online Web Spam **Detection System**∗
www.cs.sfu.ca/~jpei/publications/pagefarm-kdd09demo.pdf
File Format: PDF/Adobe Acrobat - Quick View
by B Zhou - Related articles
Ranking web pages is an essential task in **web search** and ... the quality of **information retrieval** on the web. ... ample of such off-**search engine** applications. ...

DOCODE-lite: a meta-**search engine** for document similarity retrieval ...
uchile.academia.edu/.../DOCODE-lite_a_meta-**search_engine**_for_do...
A second alternative is to submit the whole document in a **search engine**. ... for which it is highly scalable and extensible to new **information retrieval** requirements. ... the average relevance of the best response of a **Web search engine** s, and β .... DOCODE-Lite is a first approach for a document plagiarism **detection system**. ...

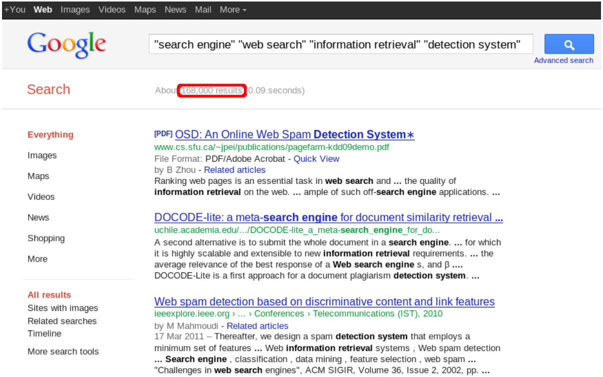
Web spam detection based on discriminative content and link features
ieeexplore.ieee.org › ... › Conferences › Telecommunications (IST), 2010
by M Mahmoudi - Related articles
17 Mar 2011 – Thereafter, we design a spam **detection system** that employs a minimum set of features ... Web **information retrieval** systems , Web spam detection ... **Search engine** , classification , data mining , feature selection , web spam ... "Challenges in **web search** engines", ACM SIGIR, Volume 36, Issue 2, 2002, pp. ...

Delete rare keywords Information lost!

information retrieval ~~text reuse~~

detection system web search ~~query formulation~~

~~capacity constrained~~ search engine



"search engine" "web search" "information retrieval" "detection system"

About 168,000 results (0.09 seconds)

[PDF] OSD: An Online Web Spam **Detection System**∗
www.cs.sfu.ca/~jpei/publications/pagefarm-kdd09demo.pdf
File Format: PDF/Adobe Acrobat - Quick View
by B Zhou - Related articles
Ranking web pages is an essential task in **web search** and **...** the quality of **information retrieval** on the web. **...** ample of such off-**search engine** applications. **...**

DOCODE-lite: a meta-**search engine** for document similarity retrieval **...**
uchile.academia.edu/.../DOCODE-lite_a_meta-**search**_**engine**_for_do...
A second alternative is to submit the whole document in a **search engine**. **...** for which it is highly scalable and extensible to new **information retrieval** requirements. **...** the average relevance of the best response of a **Web search engine** s, and β **....** DOCODE-Lite is a first approach for a document plagiarism **detection system**. **...**

Web spam detection based on discriminative content and link features
ieeexplore.ieee.org › ... › Conferences › Telecommunications (IST), 2010
by M Mahmoudi - Related articles
17 Mar 2011 – Thereafter, we design a spam **detection system** that employs a minimum set of features **...** Web **information retrieval** systems , Web spam detection **...** **Search engine** , classification , data mining , feature selection , web spam **...** "Challenges in **web search** engines", ACM SIGIR, Volume 36, Issue 2, 2002, pp. **...**

Again . . .

What query to formulate from the keywords?

Our answer . . .

Not just one query!

Our answer . . .

Not just one query!

But a set of queries!

Our answer . . .

Not just one query!

But a set of queries!

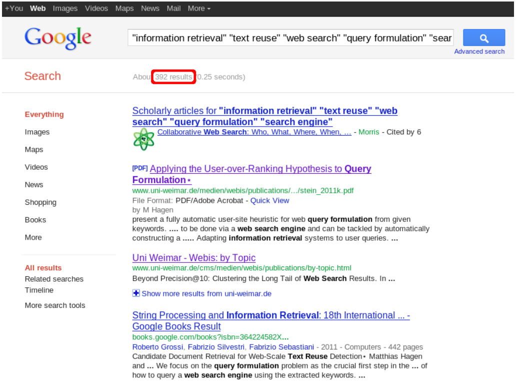
Remark: Each returning not too many results . . .

A set of queries! query 1/3

information retrieval text reuse

~~detection system~~ web search query formulation

~~capacity constrained~~ search engine



"information retrieval" "text reuse" "web search" "query formulation" "sear

About 392 results (0.25 seconds)

Scholarly articles for "information retrieval" "text reuse" "web search" "query formulation" "search engine"
Collaborative Web Search: Who, What, Where, When, ... - Morris - Cited by 6

[PDF] Applying the User-over-Ranking Hypothesis to Query Formulation
www.uni-weimar.de/medien/webis/publications/.../stein_2011k.pdf
File Format: PDF/Adobe Acrobat - Quick View
by M Hagen
present a fully automatic user-site heuristic for web **query formulation** from given keywords. .... to be done via a **web search engine** and can be tackled by automatically constructing a ..... Adapting **information retrieval** systems to user queries. ...

Uni Weimar - Webis: by Topic
www.uni-weimar.de/cms/medien/webis/publications/by-topic.html
Beyond Precision@10: Clustering the Long Tail of **Web Search** Results. In ...
Show more results from uni-weimar.de

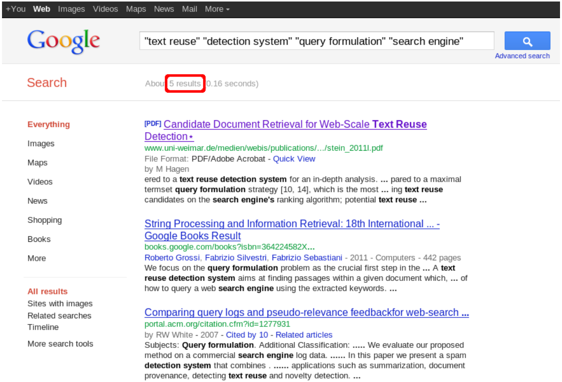
String Processing and Information Retrieval: 18th International ... - Google Books Result
books.google.com/books?isbn=364224582X...
Roberto Grossi, Fabrizio Silvestri, Fabrizio Sebastiani - 2011 - Computers - 442 pages
Candidate Document Retrieval for Web-Scale **Text Reuse** Detection Matthias Hagen and ... We focus on the **query formulation** problem as the crucial first step in the ... of how to query a **web search engine** using the extracted keywords. ...

A set of queries! query 2/3

~~information retrieval~~ text reuse

detection system ~~web search~~ query formulation

~~capacity constrained~~ search engine



"text reuse" "detection system" "query formulation" "search engine"

Abou 5 results 0.16 seconds)

[PDF] Candidate Document Retrieval for Web-Scale **Text Reuse** Detection
www.uni-weimar.de/medien/webis/publications/.../stein_2011l.pdf
File Format: PDF/Adobe Acrobat - Quick View
by M Hagen
ered to a **text reuse detection system** for an in-depth analysis. ... pared to a maximal termset **query formulation** strategy [10, 14], which is the most ... ing **text reuse** candidates on the **search engine's** ranking algorithm; potential **text reuse** ...

String Processing and Information Retrieval: 18th International ... - Google Books Result
books.google.com/books?isbn=364224582X...
Roberto Grossi, Fabrizio Silvestri, Fabrizio Sebastiani - 2011 - Computers - 442 pages
We focus on the **query formulation** problem as the crucial first step in the ... A **text reuse detection system** aims at finding passages within a given document which, ... of how to query a web **search engine** using the extracted keywords. ...

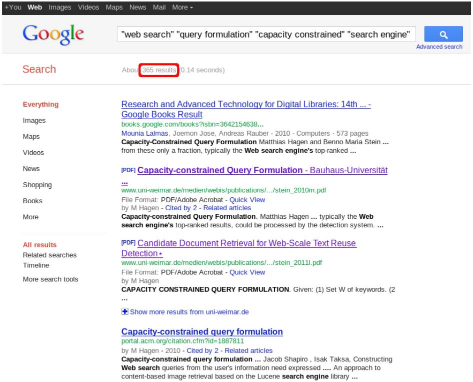
Comparing query logs and pseudo-relevance feedbackfor web-search ...
portal.acm.org/citation.cfm?id=1277931
by RW White - 2007 - Cited by 10 - Related articles
Subjects: **Query formulation**. Additional Classification: ..... We evaluate our proposed method on a commercial **search engine** log data. ...... In this paper we present a spam **detection system** that combines . ... applications such as summarization, document provenance, detecting **text reuse** and novelty detection. ...

A set of queries! query 3/3

~~information retrieval~~ ~~text reuse~~

~~detection system~~ web search query formulation

capacity constrained search engine



"web search" "query formulation" "capacity constrained" "search engine"

About 365 results (0.14 seconds)

Research and Advanced Technology for Digital Libraries: 14th ... - Google Books Result
books.google.com/books?isbn=3642154638...
Mounia Lalmas, Joemon Jose, Andreas Rauber - 2010 - Computers - 573 pages
Capacity-Constrained Query Formulation Matthias Hagen and Benno Maria Stein ... from these only a fraction, typically the **Web search engine's** top-ranked ...

[PDF] Capacity-constrained Query Formulation - Bauhaus-Universität ...
www.uni-weimar.de/medien/webis/publications/.../stein_2010m.pdf
File Format: PDF/Adobe Acrobat - Quick View
by M Hagen - Cited by 2 - Related articles
Capacity-constrained Query Formulation. Matthias Hagen ... typically the **Web search engine's** top-ranked results, could be processed by the detection system. ...

[PDF] Candidate Document Retrieval for Web-Scale Text Reuse Detection
www.uni-weimar.de/medien/webis/publications/.../stein_2011l.pdf
File Format: PDF/Adobe Acrobat - Quick View
by M Hagen
CAPACITY CONSTRAINED QUERY FORMULATION. Given: (1) Set W of keywords. (2 ...

Show more results from uni-weimar.de

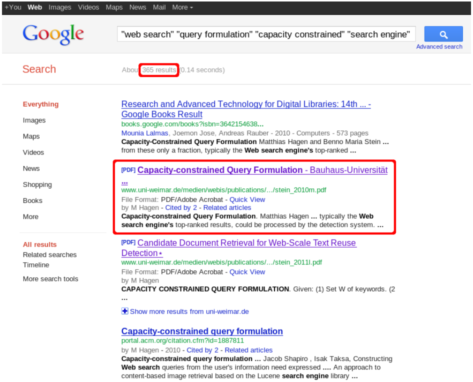
Capacity-constrained query formulation
portal.acm.org/citation.cfm?id=1887811
by M Hagen - 2010 - Cited by 2 - Related articles
Capacity-constrained query formulation ... Jacob Shapiro , Isak Taksa, Constructing **Web search** queries from the user's information need expressed .... An approach to content-based image retrieval based on the Lucene **search engine** library ...

A set of queries! query 3/3

~~information retrieval~~ ~~text reuse~~

~~detection system~~ web search query formulation

capacity constrained search engine



"web search" "query formulation" "capacity constrained" "search engine"

About 365 results (0.14 seconds)

Research and Advanced Technology for Digital Libraries: 14th ... - Google Books Result
books.google.com/books?isbn=3642154638...
Mounia Lalmas, Joemon Jose, Andreas Rauber - 2010 - Computers - 573 pages
Capacity-Constrained Query Formulation Matthias Hagen and Benno Maria Stein **...** from these only a fraction, typically the **Web search engine's** top-ranked **...**

[PDF] **Capacity-constrained Query Formulation** - Bauhaus-Universität ...
www.uni-weimar.de/medien/webis/publications/.../stein_2010m.pdf
File Format: PDF/Adobe Acrobat - Quick View
by M Hagen - Cited by 2 - Related articles
Capacity-constrained Query Formulation. Matthias Hagen **...** typically the **Web search engine's** top-ranked results, could be processed by the detection system. **...**

[PDF] Candidate Document Retrieval for Web-Scale Text Reuse Detection
www.uni-weimar.de/medien/webis/publications/.../stein_2011l.pdf
File Format: PDF/Adobe Acrobat - Quick View
by M Hagen
CAPACITY CONSTRAINED QUERY FORMULATION. Given: (1) Set W of keywords. (2 ...

Show more results from uni-weimar.de

Capacity-constrained query formulation
portal.acm.org/citation.cfm?id=1887811
by M Hagen - 2010 - Cited by 2 - Related articles
Capacity-constrained query formulation ... Jacob Shapiro , Isak Taksa, Constructing **Web search** queries from the user's information need expressed **....** An approach to content-based image retrieval based on the Lucene **search engine** library **...**

The 3 queries together . . .

Properties

- All keywords covered (similarity)
- Not too many results (≤ 1000) (capacity)
- Desired document among the results (quality)

Problem

How to automatically find such query sets?

Problem statement

Capacity Constrained Query Formulation

- Given:
  1. Set W of keywords
  2. Query interface for a web search engine
  3. Upper bound k on the number of desired results
- Find a family $\mathcal{Q} \subseteq 2^W$ of queries:
  - returning $\leq k$ results
  - covering all keywords from W.

Optimization Problem!

Minimize the number of submitted web queries to find $\mathcal{Q}$.

Problem statement

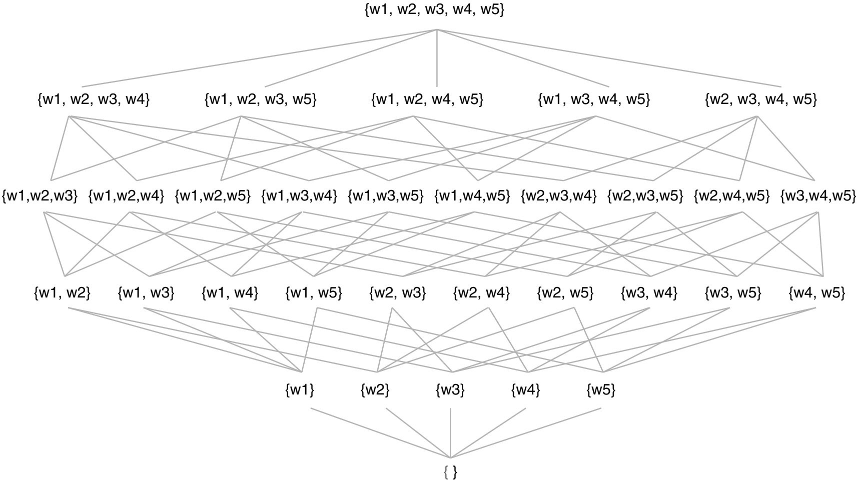
Capacity Constrained Query Formulation

- Given:
  1. Set W of keywords
  2. Query interface for a web search engine
  3. Upper bound k on the number of desired results
- Find a family $\mathcal{Q} \subseteq 2^W$ of queries:
  - returning $\leq k$ results
  - covering all keywords from W.

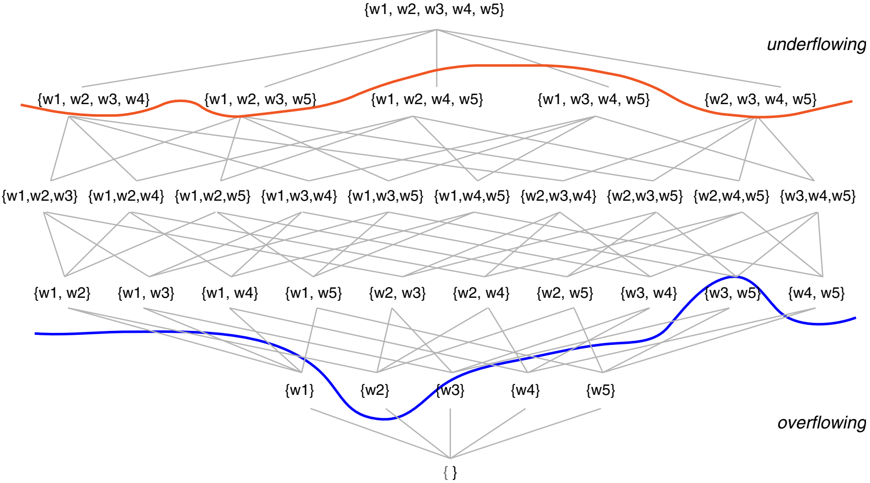
Optimization Problem!

Minimize the number of submitted web queries to find $\mathcal{Q}$.

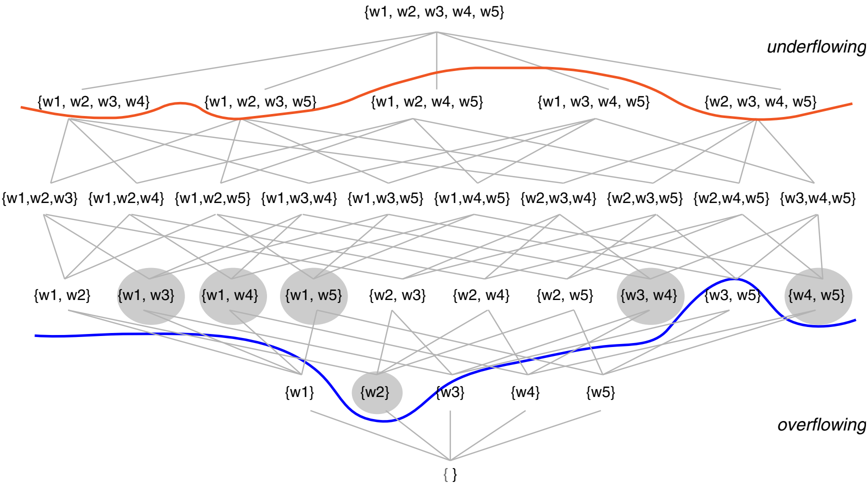
All possible queries



Queries with at most ℓ results



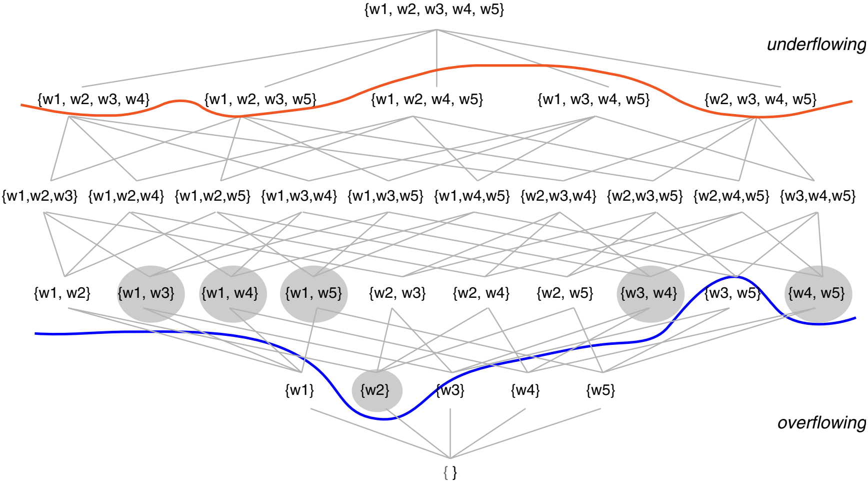
Minimal non-overflowing queries



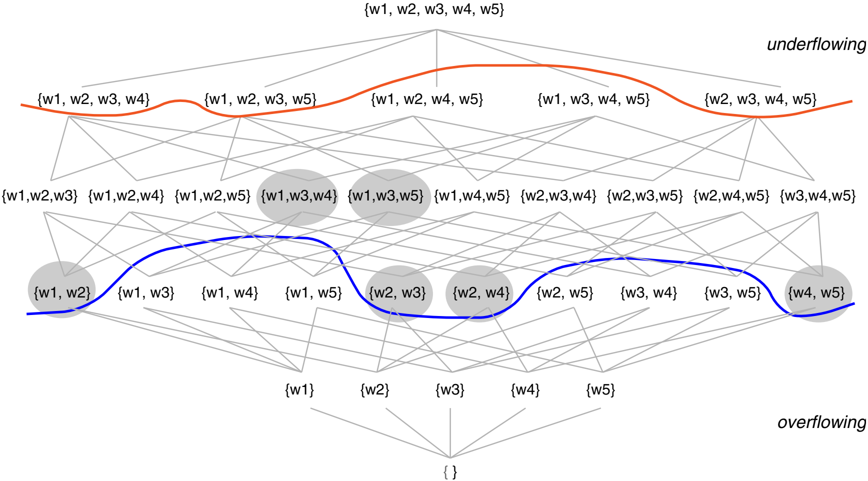
The baseline algorithm

Apriori

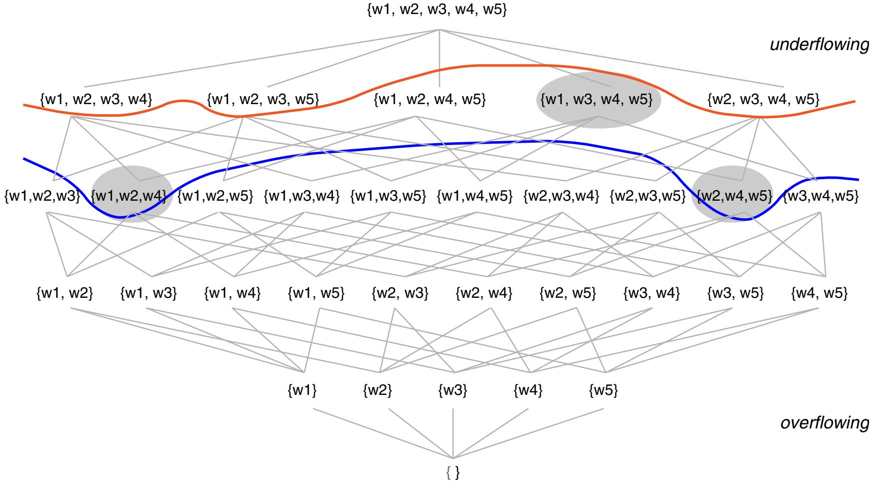
Apriori Step 1 $\ell = 1000$ 4500 results



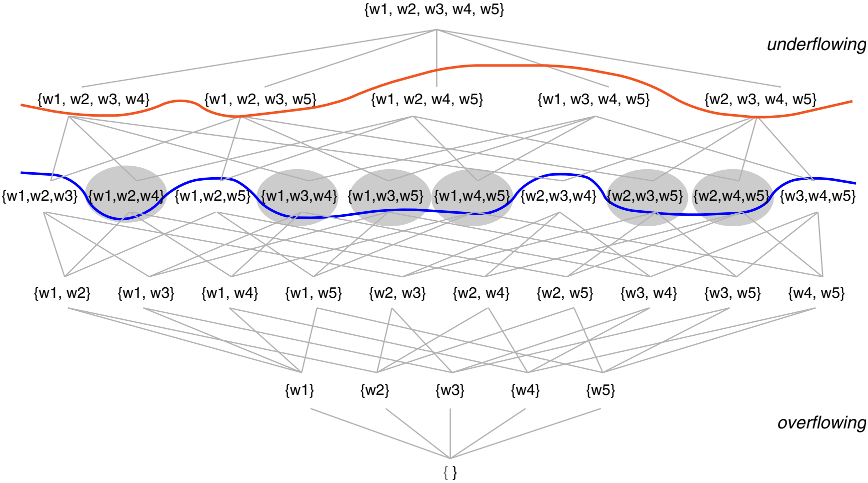
Apriori Step 2 $\ell = 500$ 2700 results



Apriori Step 3 $\ell = 250$ 380 results



Apriori Step 4 $\ell = 375$ 940 results



Baseline's Analysis

Major drawback

All intermediate queries submitted. → Bad run time!

Baseline's Analysis

Major drawback

All intermediate queries submitted. → Bad run time!

Idea

Estimate the result list length before query submission.

The improved heuristic

Apriori + estimation

Co-occurrences for estimation

Estimate: `"information retrieval" "query formulation" + "web search"`

Co-occurrences for estimation

Estimate: `"information retrieval" "query formulation"` + `"web search"`

Known: `"information retrieval" "query formulation"` 87 100 results

Co-occurrences for estimation

Estimate: `"information retrieval" "query formulation" + "web search"`

Known:

`"information retrieval" "query formulation"`	87 100 results
`"information retrieval" + "web search"`	16 % remain

Co-occurrences for estimation

Estimate: `"information retrieval" "query formulation" + "web search"`

Known:	`"information retrieval" "query formulation"`	87 100 results
	`"information retrieval" + "web search"`	16 % remain
	`"query formulation" + "web search"`	22 % remain

Co-occurrences for estimation

Estimate: `"information retrieval" "query formulation" + "web search"`

Known:		
	`"information retrieval" "query formulation"`	87 100 results
	`"information retrieval"      + "web search"`	16 % remain
	`"query formulation"          + "web search"`	22 % remain

Our estimation scheme:

$$\mathrm{avg}(16\,\%\,,\ 22\,\%) = 19\,\%$$

$$87\,100 \cdot 0.19 = 16\,500 \text{ results}$$

Co-occurrences for estimation

Estimate: `"information retrieval" "query formulation" + "web search"`

Known:	`"information retrieval" "query formulation"`	87 100 results
	`"information retrieval" + "web search"`	16 % remain
	`"query formulation" + "web search"`	22 % remain

Our estimation scheme:

$$\mathrm{avg}(16\,\%\,,\ 22\,\%) = 19\,\%$$

$$87\,100 \cdot 0.19 = 16\,500 \text{ results}$$

Control: Google 35 700 results

Co-occurrences for estimation

Estimate: `"information retrieval" "query formulation" + "web search"`

Known:	`"information retrieval" "query formulation"`	87 100 results
	`"information retrieval"       + "web search"`	16 % remain
	`"query formulation"           + "web search"`	22 % remain

Our estimation scheme:

$$\mathrm{avg}(16\,\%\,,\ 22\,\%) = 19\,\%$$

$$87\,100 \cdot 0.19 = 16\,500 \text{ results}$$

Control: Google 35 700 results

Observation

Our scheme usually underestimates the real result list length.

What about performance?

Experimental setup

Corpus

- 257 pairs of two versions of papers
- 10 keywords from more mature version

System

- Bing API as search engine
- Set $k = 1000$

Experimental setup

Corpus

- 257 pairs of two versions of papers
- 10 keywords from more mature version

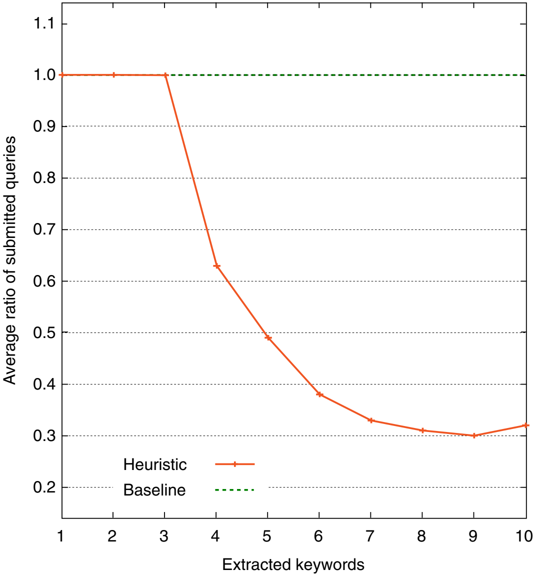
System

- Bing API as search engine
- Set $k = 1000$

Baseline vs. heuristic

Number of keywords		4	6	8	10
complete query overflows		207	146	102	81
$\mathcal{Q}$ computation possible		50	111	155	176
Avg. queries submitted	heuristic	**6.69**	**13.30**	**32.58**	**95.86**
	baseline	10.65	34.60	106.19	302.87

Baseline vs. heuristic



What about the candidate document quality?

Candidates' similarity to original document

	Approach			
	Heuristic	Frequent	Rare	Random
10 most similar doc's	0.55	0.55	**0.56**	**0.56**
100 most similar doc's	**0.39**	0.37	0.35	0.29
all retrieved doc's	**0.29**	0.25	0.22	0.21

Almost the end: The take-away messages!

What we have done

Results

- Candidate document retrieval
  - not just one query
  - set of queries → capacity
- Co-occurrence informed heuristic
- Good quality candidates

Future work

- Which approach actually finds more text reuse?

What we have (not) done

Results

- Candidate document retrieval
  - not just one query
  - set of queries → capacity
- Co-occurrence informed heuristic
- Good quality candidates

Future work

- Which approach actually finds more text reuse?

What we have (not) done

Results

- Candidate document retrieval
  - not just one query
  - set of queries → capacity
- Co-occurrence informed heuristic
- Good quality candidates

Future work

- Which approach actually finds more text reuse?

Thank you

☺