

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.

Web-based Plagiarism Analysis
Alexander Meyer, Benno Stein, Christoph Lohmann, Gernot Heiser, and Hans-Joachim Lischke
Faculty of Informatics, University of Duisburg-Essen, Essen, Germany
meyer@uni-due.de, stein@uni-due.de, lohmann@uni-due.de, heiser@uni-due.de, lischke@uni-due.de

Abstract This paper is based on a novel Web-based application for the analysis of text documents with respect to plagiarism. After first reports on experiences with classical algorithms, a new method for plagiarism analysis is introduced. After web-based algorithms for plagiarism detection are presented, the effectiveness of a similarity-based method is experimentally evaluated. The results can be compared, they are extended to deal particularly with passages using only the target document. The idea of plagiarism analysis is extended to a) when passages are copied from several documents and b) when the similarity-based method is extended to deal with multiple and consequently to identify multiple passages within a single document. Apart from the possibility to solve the original problem, the presented method can also be used to trace a source for potentially copied documents.

1 Introduction
Text reuse analysis, also called, keyword search, clustering, [1][2][3][4][5][6][7][8][9][10][11][12][13][14][15][16][17][18][19][20][21][22][23][24][25][26][27][28][29][30][31][32][33][34][35][36][37][38][39][40][41][42][43][44][45][46][47][48][49][50][51][52][53][54][55][56][57][58][59][60][61][62][63][64][65][66][67][68][69][70][71][72][73][74][75][76][77][78][79][80][81][82][83][84][85][86][87][88][89][90][91][92][93][94][95][96][97][98][99][100]

2 Plagiarism Types
Plagiarism is the use of another's ideas, information, language, or writing without the proper acknowledgment of the original source [1]. There are many types of plagiarism, including direct plagiarism, which is the copying of text from one source and presenting it as one's own work. Other types of plagiarism include paraphrasing, which is the restating of text in one's own words, and self-plagiarism, which is the reuse of one's own work in a new document. Plagiarism can be intentional or unintentional. Plagiarism is a serious offense in many fields, including academia, journalism, and the arts. Plagiarism can be detected using various methods, including manual review and computer programs.

Our Web-based plagiarism analysis application takes a complete document (text or pdf) as input (see Figure 1). Consequently, all unprocessed document content is passed to the algorithm that then is able to detect text reuse (text or pdf) in documents, language independent being a plus. Hidden and hidden passages can be marked [1] to trace one or all contents of text or document data [2].

All Query Document Processing Details
When inputs are extracted from the complete document, we employ a heuristic query generation procedure, which was first proposed by [3]. Let us describe our set of heuristic features that are extracted from a complete document. By using synonyms, acronyms, terms, and derivationally related forms, the set [4] is extracted through a search [5] utilizing groups of words are identified by creating statistical knowledge about the text [6] and their relations, as well as selecting or creating words, adding the set [7] [8]. Then, a comparison of queries is generated (and passed to search engines).

This solution step is extended by quantitative relevance analysis. Depending on the number of used documents (one or two "word"/ queries are generated. Note that each a context can be realized by a heuristic analysis of the set [9], which contains word groups and word frequency classes [10]. The result of this step is a complete document collection [11][12][13][14][15][16][17][18][19][20][21][22][23][24][25][26][27][28][29][30][31][32][33][34][35][36][37][38][39][40][41][42][43][44][45][46][47][48][49][50][51][52][53][54][55][56][57][58][59][60][61][62][63][64][65][66][67][68][69][70][71][72][73][74][75][76][77][78][79][80][81][82][83][84][85][86][87][88][89][90][91][92][93][94][95][96][97][98][99][100]

3 Plagiarism Analysis
An original document, a document may be plagiarized in different ways. Consequently, several indicators can be proposed to detect plagiarism. An adoption of indicators that are given in [1] is as follows.

- (1) Copied text. If text comes from a source that is known and it is not cited properly then this is an instance of plagiarism.
- (2) Misquoting. If the reference to documents overlap significantly, the bibliography and other parts may be copied. A changing citing style may be a sign for theft.
- (3) Change in writing style. A copied change in the author's style may appear paragraphs or sections with, e.g., different adjectives and adjectives style, unbalanced word style, better and better processes.
- (4) Change in structure. In copy and paste plagiarism cases the formatting of the original document is inherited by partial paragraphs, especially when content is copied from literature to text processing programs.
- (5) Textual performance. If the flow of expressions throughout a document is consequently broken then the document may be a "stolen plagiarism", i.e. a compilation of different sources.



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

by Martin Peckler, Daily Mirror 10/10/2011

Recommended 10 Saves 10 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 98 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 1859.

The peak of the 325ft 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 Parliament House, used as MPs' offices – in 5,000 years.



Leaning tower of London? Big Ben is tilting

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



LONDON | Tue Oct 11, 2011 6:13AM 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 Member'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 1859.

MOST POPULAR

READ

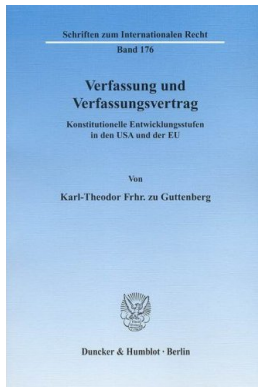
- 1 AOL CEO pitches investors an 'ultra-low' stock price 08 Oct 2011
- 2 Alleged Iran plot 'blows up' ex-collector: Clinton 08 Oct 2011
- 3 Obama's former CEO claims U.S. walls are 'ineffective' 08 Oct 2011
- 4 Alabama immigration law denied, upheld on same day again 08 Oct 2011
- 5 U.S. not bigger than thought, and very hungry 08 Oct 2011

ENGAGED

- 171 California governor signs controversial 'Dream Act' 08 Oct 2011
- 128 Hank Williams Jr. lashes out at media in new song 08 Oct 2011
- 116 Insight: Occupy Wall St, the start of a new protest era? 08 Oct 2011



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



60% of dissertation plagiarized

Candidate Document Retrieval for Web-Scale Text Reuse Detection*

Matthias Hagen and Benno Stein

Faculty of Media
Bielefeld University
-first name-@uni-bielefeld.de

Abstract Given a document d , the task of text reuse detection is to find those passages in d which are identical or paraphrased from 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-termed query formulation strategy [10, 14], which is the most scalable 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, 9].

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 wherein 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 "preexisting" 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

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 - first initial

Abstract Given a document d , the task of text reuse detection is to find those passages in d which is 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 scalable 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, 9].

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 wherein text reuse detection addresses a broader spectrum that also covers phrases 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.

One 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].

Capacity-constrained Query Formulation

Matthias Hagen and Benno Maria Stein

Faculty of Media
Bauhaus University Weimar, Germany

Abstract Given a set of keywords, we analyze how Web queries with these phrases can be formed that, taken together, return a specified number of hits. The core case of this problem is a plagiarism detection system that searches the Web for potentially plagiarized passages in a given suggestion document. For the query formulation problem we develop a heuristic search strategy based on co-occurrence probabilities. Compared to the maximal termset strategy [1], which can be considered as the most scalable 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 suggestion document can be found via direct comparisons against potential source documents. Today's typical source of plagiarisms is the Web, which obviously contains too many documents for direct comparisons. The straightforward solution is to extract keywords from the suggestion document and to retrieve a installable number of documents containing these phrases. These documents are considered as the best potential sources of plagiarisms 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 is able to direct comparisons is constrained by some processing capacity k . If all the extracted keywords (usually about 100 from the suggestion document) are substituted as one single Web query, probably too few documents are returned with respect to k . Similarly, queries consisting 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 plagiarisms 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-kisses-donkey hypothesis* or, more formally, *user-over-loading 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-loading hypothesis the CAPACITY-CONSTRAINED QUERY FORMULATION problem analyzed in this paper is defined as follows. Given (1) a set W of keywords, (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 $Q \subseteq 2^W$ 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 Q , and we focus on the following optimization problem from the detection system's perspective: What strategy

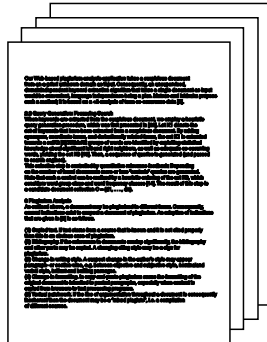
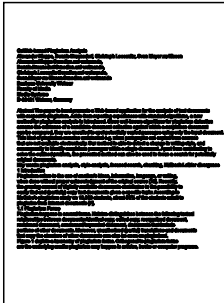
SPIRE 2011 full paper

ECJL 2010 poster

Text reuse detection

Given “suspicious” document

Step 1: Find a set of candidate documents

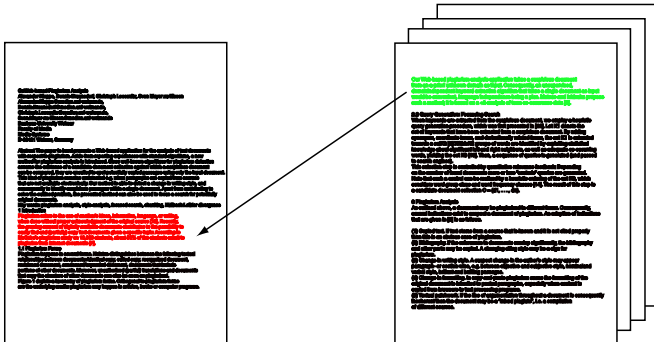


Text reuse detection

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

Observations

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

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

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?

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 d' which are identical or paraphrased from 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 filtered in 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-termed 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
christ.nano@clat.nano-stf.uni-weimar.de

Abstract Given a document d , the task of text reuse detection is to find those passages in d which are identical or paraphrased from also appearing 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 filtered in 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-termed query formulation strategy [10, 14], which is the most sensible non-heuristic baseline. In our experiments, we see 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 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 occurrence probabilities. Compared to the maximal-termed strategy [5], which can be considered as the most sensible non-heuristic baseline, our proposed strategy can 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. Today's 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/retrieval

detectable/system

web/search

query/algorithm

detectable/algorithm

search/algorithm

The screenshot shows a Google search interface. At the top, there are navigation links: '+ You', 'Web', 'Images', 'Videos', 'Maps', 'News', 'Mail', and 'More'. The Google logo is on the left, and the search bar contains the text '"information retrieval"'. To the right of the search bar is a blue button with a magnifying glass icon and the text 'Advanced search'. Below the search bar, the results are displayed. The first result is 'Information retrieval - Wikipedia, the free encyclopedia' with a URL 'en.wikipedia.org/wiki/Information_retrieval'. The second result is 'Introduction to Information Retrieval' with a URL 'nlp.stanford.edu/IR-book/'. The third result is 'Information Retrieval' with a URL 'www.dcs.gla.ac.uk/Kelth/Preface.html'. The search results are organized into a table with a left column for categories and a right column for search results.

Search	About (19,900,000 results) (0.07 seconds)
Everything	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 ...
Images	
Maps	
Videos	
News	Introduction to Information Retrieval nlp.stanford.edu/IR-book/ 25+ items – Introduction to Information Retrieval . This is the companion ... <ul style="list-style-type: none">• Front matter (incl. table of notations – pdf• 02 – The term vocabulary & postings lists• 03 – Dictionaries and tolerant retrieval
Shopping	
Blogs	
Books	Slides - Irbook - Text classification and Naive Bayes - Exercises
Discussions	Information Retrieval www.dcs.gla.ac.uk/Kelth/Preface.html An online book by C. J. van Rijsbergen, University of Glasgow.
More	

Single keyword queries?

detection system

Handwritten-style text representing the words "detection system" and "search results" in various orientations and colors.


The screenshot shows a Google search interface. At the top, there are navigation links: "+You Web Images Videos Maps News Mail More -". The Google logo is on the left, and the search input field contains the text "detection system". To the right of the input field is a blue search button with a magnifying glass icon and the text "Advanced search". Below the search bar, the results are displayed. The word "Search" is in red. To the right, it says "About 18,500,000 results (0.16 seconds)", with "18,500,000 results" highlighted in a red box. On the left side, there are filters for "Everything", "Images", "Maps", "Videos", "News", "Shopping", and "More". Under "Everything", there are several search results. The first result is "Intrusion detection system - Wikipedia, the free encyclopedia" with a link to en.wikipedia.org/wiki/Intrusion_detection_system. The second result is "An intrusion detection system (IDS) is a device or software application that monitors network and/or system activities for malicious activities or policy violations ..." with a link to "Snort - Network intrusion detection system - Host-based intrusion detection system". The third result is "What is intrusion detection system? - A Word Definition From the ..." with a link to www.webopedia.com/TERM/I/intrusion_detection_system.html. The fourth result is "Intrusion Detection System" with a link to www.intrusiondetectionsystem.org/. The fifth result is "Snort :: Home Page" with a link to www.snort.org/. The sixth result is "Snort® is an open source network intrusion prevention and detection system (IDS/IPS) developed by Sourcefire. Combining the benefits of signature, protocol, ...".

detection system

http://www.ietf.org/rfc/rfc2818.html *text/news*
news/search *query/attack*
http://blogs.ubuntu.com *search/elite*

The screenshot shows a Google search interface. At the top, there are navigation links for Web, Images, Videos, Maps, News, Mail, and More. The search bar contains the text "detection system" and a search button with a magnifying glass icon. Below the search bar, the text "Search" is displayed in red, followed by "About 18,500,000 results (0.16 seconds)". The results are categorized into "Everything" and "Any time". Under "Everything", there are links to Wikipedia, a definition from webopedia, and Snort's home page. Under "Any time", there are links for Past 24 hours, Past 2 days, Past week, Past month, and Past year.

+You **Web** Images Videos Maps News Mail More -

Google "detection system"  Advanced search

Search About **18,500,000 results** (0.16 seconds)

Everything

- [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.intrusiondetectionsystem.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, ...

Any time

- Past hour
- Past 24 hours
- Past 2 days
- Past week
- Past month
- Past year

information retrieval text reuse
detection system web search query formulation
capacity constrained search engine

The screenshot shows a Google search interface. At the top, navigation links include '+You', 'Web', 'Images', 'Videos', 'Maps', 'News', 'Mail', and 'More'. The Google logo is on the left, and the search bar contains the query: "information retrieval" "text reuse" "detection system" "web search" "query formulation" "capacity constrained" "search engine". A blue search button with a magnifying glass icon is on the right, with 'Advanced search' text below it. Below the search bar, the word 'Search' is written in red. On the left side, there is a vertical menu with 'Everything' in red, followed by 'Images', 'Maps', 'Videos', 'News', 'Shopping', and 'More'. The main content area shows the search results: 'Your search - "information retrieval" "text reuse" "detection system" "web search" "query formulation ... - did not match any documents.' The phrase 'not match any' is highlighted with a red box. Below this, there is a 'Suggestions:' section with a bulleted list: 'Make sure all words are spelled correctly.', 'Try different keywords.', 'Try more general keywords.', and 'Try fewer keywords.' At the bottom, there are links for 'Show search tools', 'Search Help', and 'Give us feedback'. The footer contains links for 'Google Home', 'Advertising Programs', 'Business Solutions', 'Privacy', and 'About Google'.

information retrieval text reuse
detection system web search query formulation
capacity constrained search engine

The screenshot shows a Google search interface. At the top, navigation links include '+You', 'Web', 'Images', 'Videos', 'Maps', 'News', 'Mail', and 'More'. The Google logo is on the left, and the search bar contains the query: "information retrieval" "text reuse" "detection system" "web search" "query formulation" "capacity constrained" "search engine". A blue search button with a magnifying glass icon and the text 'Advanced search' is on the right. Below the search bar, the word 'Search' is displayed in red. On the left side, there is a vertical menu with 'Everything' in red, followed by 'Images', 'Maps', 'Videos', 'News', 'Shopping', and 'More'. The main content area shows the search results: 'Your search - "information retrieval" "text reuse" "detection system" "web search" "query formulation ... - did not match any documents.' The phrase 'not match any' is highlighted with a red box. Below this, 'Suggestions:' are listed with three bullet points: 'Make sure all words are spelled correctly.', 'Try different keywords.', and 'Try more general keywords.' At the bottom, there are links for 'Show search tools', 'Search Help', 'Give us feedback', 'Google Home', 'Advertising Programs', 'Business Solutions', 'Privacy', and 'About Google'.

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

information retrieval

~~information retrieval~~

detection system web search

~~detection system~~ ~~web search~~

~~information retrieval~~

search engine

Search About **118,000 results** (0.09 seconds)

Everything

[PDF QSD: An Online Web Spam Detection System*](#)

Images

www.cs.sfu.ca/~jpei/publications/pagefarm-kdd09demo.pdf

Maps

File Format: PDF/Adobe Acrobat - Quick View

Videos

by B Zhou - Related articles

News

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. ...

Shopping

[DOCODE-lite: a meta-search engine for document similarity retrieval ...](#)

More

uchile.academia.edu/.../DOCODE-lite_a_meta-search_engine_for_do...

All results

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_i and β ...

Sites with images

DOCODE-Lite is a first approach for a document plagiarism **detection system**. ...

Related searches

[Web spam detection based on discriminative content and link features](#)

Timeline

ieeexplore.ieee.org > ... > Conferences > Telecommunications (IST), 2010

More search tools

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. ...

information retrieval

~~information retrieval~~

detection system web search

~~detection system web search~~~~information retrieval~~

search engine

+You **Web** Images Videos Maps News Mail More +

Google Advanced search

Search About 118,000 results (0.09 seconds)

Everything

Images

Maps

Videos

News

Shopping

More

All results

Sites with images

Related searches

Timeline

More search tools

[PDF QSD: An Online Web Spam Detection System*](#)www.cs.sfu.ca/~jpei/publications/pagefarm-kdd09demo.pdfFile 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_i 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 \(IST\)](#), 2010by 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. ...

What query to formulate from the keywords?

Not just one query!

Not just one query!

But a set of queries!

Not just one query!

But a set of queries!

Remark: Each returning not too many results ...

information retrieval

text reuse

d/dt/e/ctt/l/bn//stj/s/tem

web search

query formulation

d/dt/e/ctt/h//b/bk/s/t/h/l/d/d

search engine

The screenshot shows a Google search interface with the following elements:

- Navigation bar: +You Web Images Videos Maps News Mail More +
- Search bar: "information retrieval" "text reuse" "web search" "query formulation" "search engine" [Search]
- Advanced search link
- Search results summary: Search About (292 results) (0.25 seconds)
- Left sidebar:
 - Everything
 - Images
 - Maps
 - Videos
 - News
 - Shopping
 - Books
 - More
 - All results
 - Related searches
 - Timeline
 - More search tools
- Main results:
 - 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/webs/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 - Webs; by Topic](#)
 - [www.uni-weimar.de/cms/medien/webs/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 Stivesti, 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. ...

~~http://www.vda/teht/lehd/~~ text reuse
 detection system ~~hd/seedch~~ query formulation
~~http://ch/h/bohs/lehd~~ search engine

The screenshot shows a Google search interface with the following elements:

- Navigation bar: +You Web Images Videos Maps News Mail More -
- Search bar: Google logo, search input field containing "text reuse" "detection system" "query formulation" "search engine", and a search button.
- Search results summary: Search (5 results) (16 seconds)
- Left sidebar:
 - Everything
 - Images
 - Maps
 - Videos
 - News
 - Shopping
 - Books
 - More
 - All results
 - Sites with images
 - Related searches
 - Timeline
 - More search tools
- Main results area:
 - PDF Candidate Document Retrieval for Web-Scale Text Reuse Detection**
 - www.uni-weimar.de/medien/webis/publications/...stein_2011.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 feedback for 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. ...

*http://www.google.de/websearch**http://www.google.de**http://www.google.de/websearch*

web search

query formulation

capacity constrained

search engine

The screenshot shows a Google search interface with the search bar containing the query "web search" "query formulation" "capacity constrained" "search engine". The search bar is highlighted with a blue border. Below the search bar, the search results are displayed. The first result is a PDF document titled "Capacity-constrained Query Formulation" by Matthias Hagen and Benno Stein, published in 2010. The document is available as a PDF/Adobe Acrobat file. The search results are displayed in a list format with a sidebar on the left containing navigation options like "Everything", "Images", "Maps", "Videos", "News", "Shopping", "Books", "More", "All results", "Related searches", "Timeline", and "More search tools". The search results are also displayed in a list format with a sidebar on the left containing navigation options like "Everything", "Images", "Maps", "Videos", "News", "Shopping", "Books", "More", "All results", "Related searches", "Timeline", and "More search tools".

Search About **885 results** (0.14 seconds)

Everything

Images

Maps

Videos

News

Shopping

Books

More

All results

Related searches

Timeline

More search tools

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

PDF Capacity-constrained Query Formulation - Bauhaus-Universität
as
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_2011f.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 ...

~~http://www.volltext.de/volltext/~~ ~~http://www.volltext.de/volltext/~~
~~http://www.volltext.de/volltext/~~ web search query formulation
 capacity constrained search engine

The screenshot shows a Google search interface with the following elements:

- Navigation bar: +You Web Images Videos Maps News Mail More +
- Search bar: Google logo, search input containing "web search" "query formulation" "capacity constrained" "search engine", and an Advanced search button.
- Search results summary: Search About 300 results (0.14 seconds)
- Filters: Everything, Images, Maps, Videos, News, Shopping, Books, More
- Search results list:
 - Result 1: [Research and Advanced Technology for Digital Libraries: 14th ... - Google Books Result](#)
books.google.com/books?isbn=3642154638...
Mourika Lalmas, Joemon Jose, Andreas Rauber - 2010 - Computers - 573 pages
Capacity-Constrained Query Formulation Matthias Hagen and Benno Marta Stein ... from these only a fraction, typically the **Web search engine's** top-ranked ...
 - Result 2: [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. ...
 - Result 3: [PDF Candidate Document Retrieval for Web-Scale Text Reuse Detection](#)
www.uni-weimar.de/medien/webis/publications/.../stein_2011f.pdf
File Format: PDF/Adobe Acrobat - Quick View
by M Hagen
CAPACITY CONSTRAINED QUERY FORMULATION. Given: (1) Set W of keywords. (2) ...
 - Result 4: [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?

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 $Q \subseteq 2^W$ of queries:
 - returning $\leq k$ results
 - covering all keywords from W .

Optimization Problem!

Minimize the number of submitted web queries to find Q .

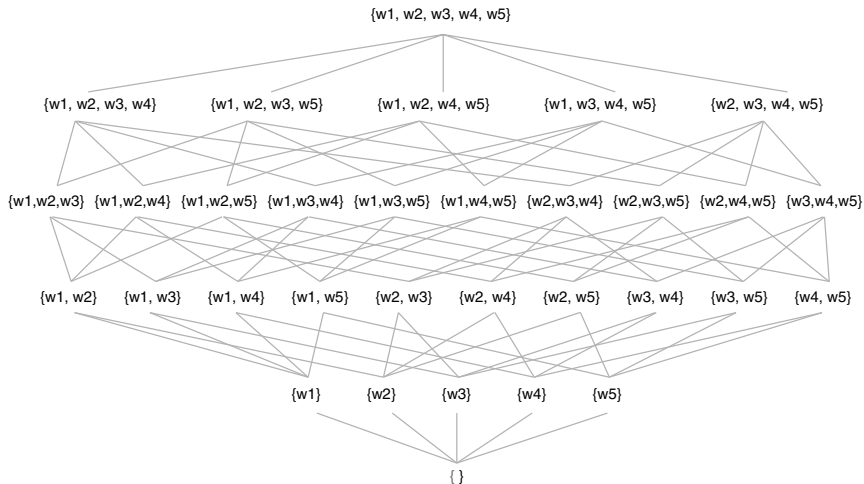
CAPACITY CONSTRAINED QUERY FORMULATION

- 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 $Q \subseteq 2^W$ of queries:
 - returning $\leq k$ results
 - covering all keywords from W .

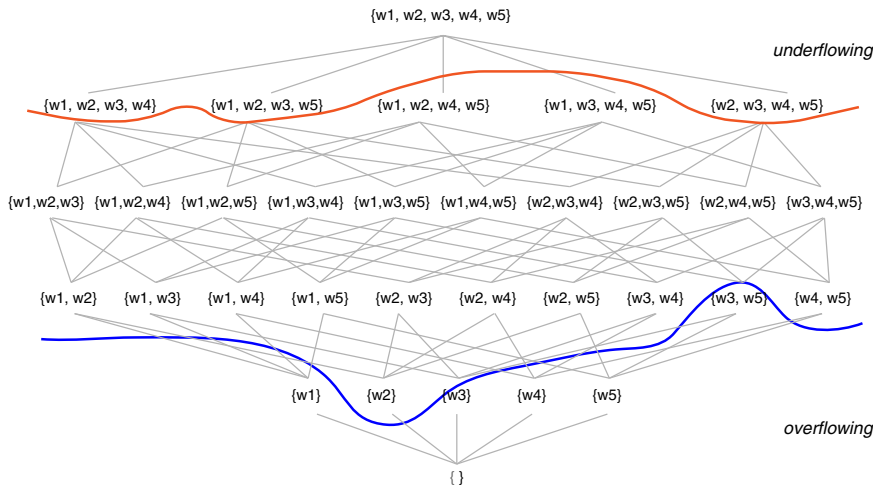
Optimization Problem!

Minimize the number of submitted web queries to find Q .

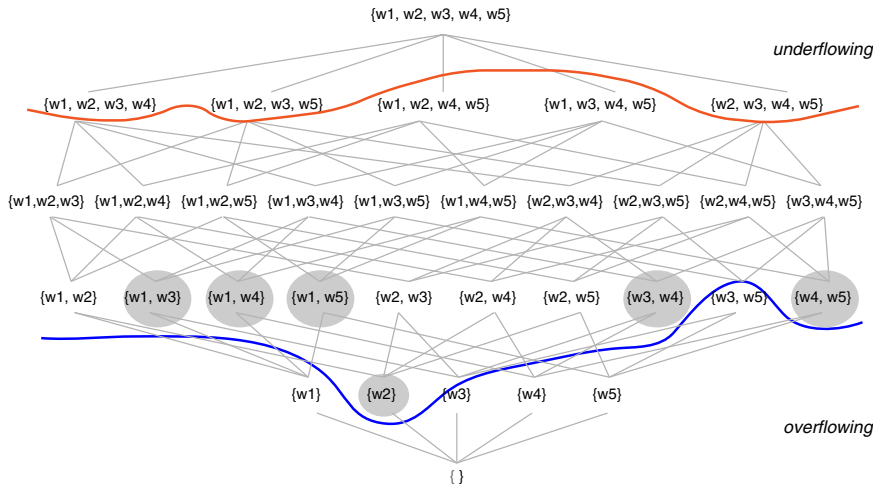
All possible queries



Queries with at most ℓ results

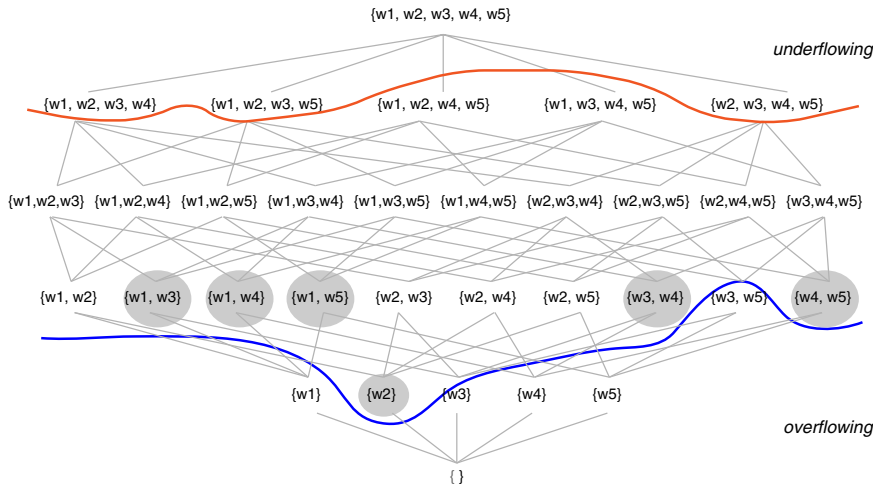


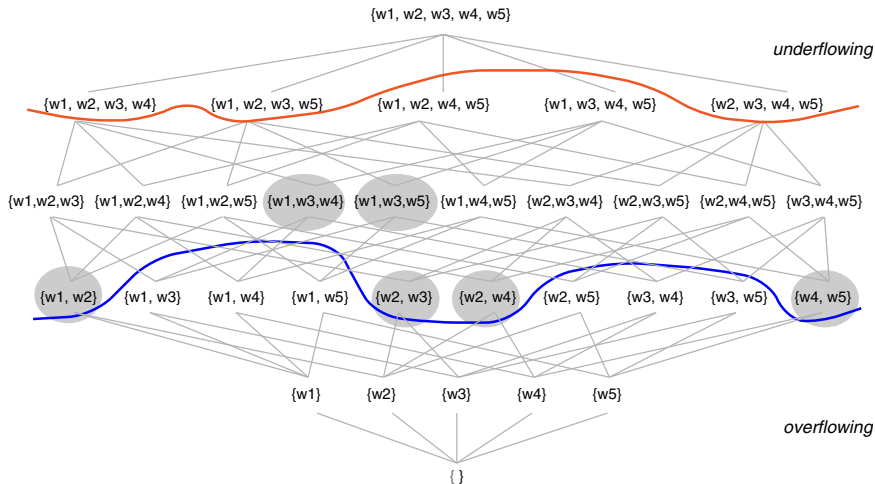
Minimal non-overflowing queries

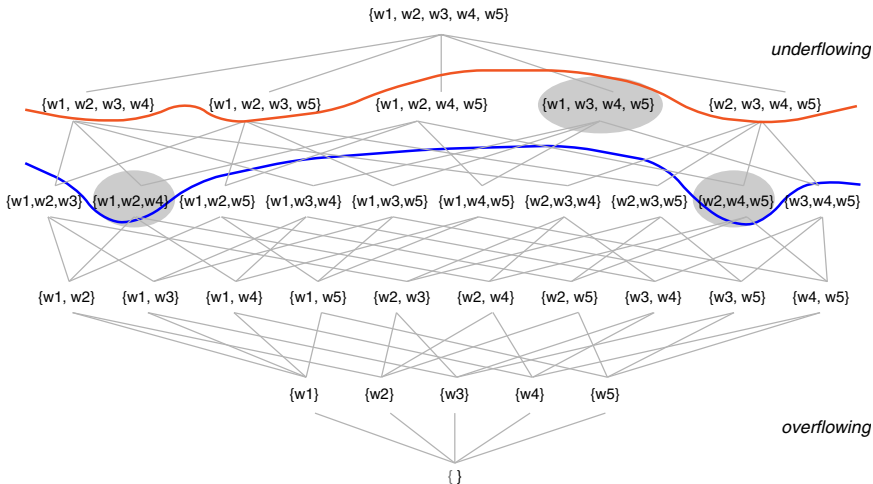


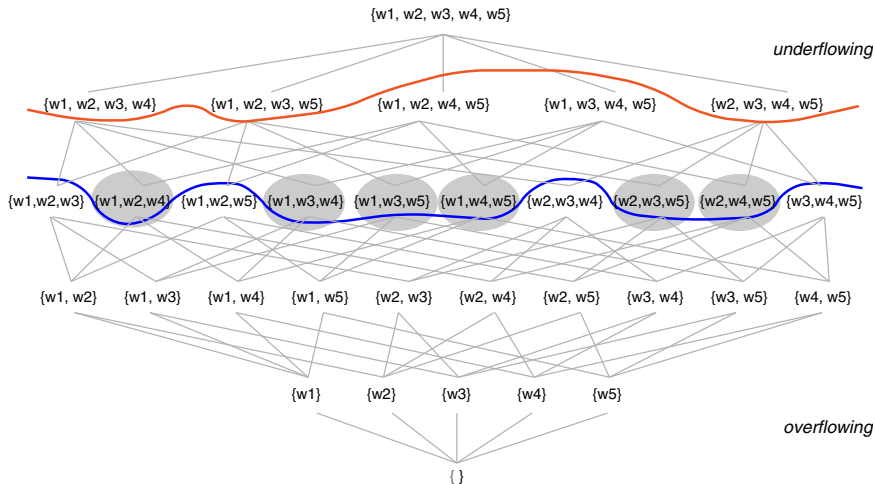
The baseline algorithm

Apriori









Major drawback

All intermediate queries submitted. → Bad run time!

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: $\text{avg}(16\% , 22\%) = 19\%$
 $87\,100 \cdot 0.19 = 16\,500$ 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: $\text{avg}(16\% , 22\%) = 19\%$
 $87\,100 \cdot 0.19 = 16\,500$ results

Control:  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: $\text{avg}(16\%, 22\%) = 19\%$
 $87\,100 \cdot 0.19 = 16\,500$ results

Control:  35 700 results

Observation

Our scheme usually underestimates the real result list length.

What about performance?

Corpus

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

System

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

Corpus

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

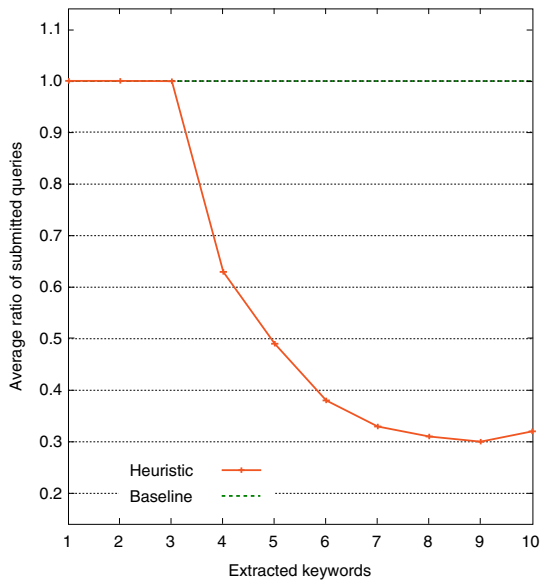
System

- 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
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!

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
