- Argumentation Analysis in Newspaper Articles
- Morning Morality
- The Super-document
- Netspeak Query Log Analysis
- Informative Linguistic Knowledge Extraction from Wikipedia
- Elastic Search and the Clueweb
- Passphone Protocol Analysis with Avispa
- Beta Web
- SimHash as a Service: Scaling Near-Duplicate Detection
- One Class Classification of Vandalism in the Wikipedia
Modeling Information Extraction Problems using Argumentation Theory

Speakers:
Philip Drewes
Jonas Köhler
Motivation:

○ Opinion mining
○ Summaries of large texts
○ Rating the validity of arguments in texts
○ Search for arguments for a given hypothesis

⇒ We want to have a computable model of argumentation for human language.
A computational model of argumentation

**nodes:** argumentative units
(Claims, Premises)

**arcs:** relations between arguments
(Attacks, Supports, ...)

**Questions:**
- When do arguments contradict?
- How are arguments related?
- What are important arguments?

**directed graph**

- We should open our borders!
- Open borders will be a threat to inner security.
- Our economy will benefit from new workers.
- Studies show there is no correlation between immigration and crime rate.
A computational model of argumentation

Searching for arguments involves the task of detecting them

Classification:

Is a part of a text an argumentative unit? ⇒ binary { yes, no }

What type of argumentative unit? ⇒ nominal { claim, premise, ... }

Are two argumentative units related? ⇒ binary { yes, no }

What type of relation is it? ⇒ nominal { attack, support, ... }

⇒ Supervised learning problem

⇒ which features?
A computational model of argumentation

Features (mostly NLP based):

Lexical: number of punctuation marks in a part of text

Syntactic: depth of the parse tree (linguistics)

Indicators: are discourse marker present?

Contextual: number of sub clauses in the sentences around the part of interest

Heavy use of the Stanford NLP Java library:

⇒ training data?

⇒ human annotation!
Creating a corpus for an argument classifier

Annotation:

Humans will annotate argumentative texts by hand.

The texts are taken from online newspapers (opinion section).

The tool for annotation is web-based. The annotations are saved to XML files.

Question:

Don’t we need 1000s of annotations?

Who will do all this work?

⇒ Crowdsourcing!
Outlook:

What we have done so far:

- Implementing a classification framework, which is
  - Calculating the feature vectors
  - Reproducing the state of the art in classification
    - Stab et al.\(^1\) achieve \(~72\%\) precision on an essay corpus
    - We are able to achieve \(~68\%\)
- Gathering the text data (automated web scraping)
- Designing the annotation job for the digital crowd.

Outlook:
What we will do until February 2015 / What may come in the long term

- Let the crowd annotate our texts and build the training corpus
- Add additional features and improve the classification
  - Extend the model? Refine the classification?
- Analyze the data
  - Which questions may arise?
- Search for argument components
  - Only possible if there is a good model + classification
Thank you for your listening!

Questions?
Morning Morality on the Web

Webis presentation

2014-12-18
Morning Morality on the Web

Project foundation and discussion starter:


Content:
- People's ethical behaviour is changing throughout the day.
- There is a "self-regulatory" resource, which depletes the longer someone is behaving good.
- Therefore, a person is more likely to cheat and lie in the afternoon or evening than in the morning.
Is such a phenomenon measurable on the Web?

• In an effort to show such an effect, Wikipedia-Vandalism cases were analyzed.

What is Wikipedia-Vandalism?

Inappropriate change, addition or removal of Wikipedia content, like adding irrelevant, abusive words, deleting pages or purposely adding false information.
How to get Wikipedia-Vandalism data?

- Scan through the history of edits for a Simple Vandalism Pattern.
- A revert back to a revision before an edit (V) is most often a case of vandalism.
- Detection is done by users and bots.
Morning Morality on the Web

previous work

Finding the „Morning Morality Effect“ in Wikipedia-Vandalism data.

Work of the previous project group:

- Analyzing correlations between local time and vandalism.
- Geolocation of vandal - IP addresses for local edit time.
Morning Morality on the Web

Finding more correlation between bad behaviour on the Web and exogenous/external factors, e.g., weather, time and region.

What we have done so far/are working on:

• Geolocate the given vandalism and normal edit dumps of the United States for 2013.
• Correlated them with the NOAA National Weather Service data (hourly weather data from 1,700 weather stations in the US over the last 15 years).
Morning Morality on the Web

current work

Early data -
Work still in progress:

US Wiki Edits 2013

Relative frequency of vandalism edits
• Analyze data for different Climate Zones and weather effects like rain and snow.

• Changing vandalism frequency in correlation with weather over time, e.g., annual and monthly time periods

• Different locations: comparisons of different states, rural and metropolitan areas.
The Super Document

A Result Presentation Paradigm for Exploratory Search Tasks

Participants:
Kevin Reinartz, Janek Bevendorff,
Kristof Komlossy, Carsten Tetens, Sebastian Gottschlich

Tim Gollub    Michael Völske    Benno Stein

Web Technology & Information Systems
Bauhaus-Universität Weimar
Winter Term 2014/15
Given the relevant documents for a query, how to present them to the user?
Weimar

1. __________________________________________

2. __________________________________________

3. __________________________________________

... 

- Compile a list of document descriptions linking to the original resources.
- Order based on the likelihood that a document contains relevant information.
Project: SuperSERP

General

- Alternative result presentation paradigms for open or undirected informational queries

- General Approach: Increase accessibility of resources in the limit of a search result list.

- Observation: An effective domain independent paradigm is hard to find.

- We concentrate on two applications:
  - Related Work Search
  - City Search
Application 1: Related Work Search

Current State

- LUCENE Index of webis-csp corpus (approx. 177,000 papers)
- Keyphrase extraction (KpMinerExtractor from aitools)
- MUSTACHE Template-Engine for search result presentation
- Search result based on keyphrases (currently)
Application 1: Related Work Search

Current user interface

**SuperSERP**

search engine

```
Search for 'tree'
```

**Outline**

1. TREE
2. ALGORITHM
3. DATA
4. NODE
5. PROBLEM
6. PROPOSED
7. STRUCTURE
8. TREES
9. NODES
10. R-TREE

**TREE**

A linear time algorithm for constructing tree 4-spanner in 2-trees

A spanning tree $T$ of a graph $G$ is said to be a tree $t$-spanner if the distance between any two vertices in $T$ is at most $t$ times their distance in $G$. A graph that has a tree $t$-spanner is called a tree $t$-spanner admissible graph. It has been shown in [3] that the problem of recognizing whether a graph admits a tree $t$-spanner is NP-complete for $t \geq 4$. In this paper, we present a linear time algor... read more

**Training tree transducers**

Many probabilistic models for natural language are now written in terms of hierarchical tree structure. Tree-based modeling still lacks many of the standard tools taken for granted in (finite-state) string-based modeling. The theory of tree transducer automata provides a possible framework to draw on, as it has been worked out in an extensive literature. We motivate the use of tree transducers... read more
Application 1: Related Work Search

Future Work

- Improved user interaction:
  - Query by document
  - Manual “topic points”
  - Quality of clustering statistics

- Topic Model for Indexing & keyquery compositing

- Efficient clustering algorithm for outline generating
Application 2: City Search

Current State

- collected Google Places
- using Bigdata as triple store (replacing Fuseki)
- read Google Places as RDF triples into triple store
- generated random people at random locations

CityBricks:

- each place is a brick
- sorted from north to south
- highlight on search & similarity
Application 2: City Search

CityBricks

Name: Zum Falken Weimar
Location: Trierer St 7
Type: restaurant
Application 2: City Search

CityTales

- take the user on a journey through the city
- create a mashup using content & statistics
- streets from a city + Random users and locations
- from various sources (Google Places, Flickr ...)
Application 2: City Search

Future Work

- Improve storytelling, infographic inspired UI
- Add sources like news, official statistics, social network, reviews
- Use focused crawling (Heritrix) to obtain web pages related to Weimar.
Netspeak Query Log Analysis
Amir Othman

cvs:
code-in-progress/webisstud/wstud-netspeak-analysis
code-in-progress/webisstud/wstud-netspeak-analysis-query-detection
code-in-progress/webisstud/wstud-netspeak-analysis-query-browser
data-in-progress/wstud-netspeak-analysis
NetSpeak

- Service to check usage of words
- ~2000 Users a month
- Log from March 2009 to February 2014
Query Detection

• Decision Tree, using log from 100 different IPs as groundtruth

• Features: overlapping characters, term overlap, character Jaccard coefficient, trigram character cosine similarity, Levenshtein distance, timegap
Netspeak Query Log Browser

- Facilitate analysis – added visualizations and interlinking
- Exploring
- Add Notes
<table>
<thead>
<tr>
<th>Query ID</th>
<th>User ID</th>
<th>Time Stamp</th>
<th>Interaction(s)</th>
<th>Number of Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>22247</td>
<td>1</td>
<td>2011-05-27 08:25:09</td>
<td>i, n, i need t,</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:10</td>
<td>i need t</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:20</td>
<td>i need</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:21</td>
<td>i need t</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:26</td>
<td>i need</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:28</td>
<td>i</td>
<td></td>
</tr>
<tr>
<td>22248</td>
<td>1</td>
<td>2011-05-27 08:25:36</td>
<td>i, ir, irt</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:36</td>
<td>i</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:37</td>
<td>ir</td>
<td></td>
</tr>
<tr>
<td>22249</td>
<td>1</td>
<td>2011-05-27 08:25:42</td>
<td>i, lets, lets mee</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:44</td>
<td>lets meet</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:46</td>
<td>lets meet o</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:47</td>
<td>lets meet on</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:51</td>
<td>lets meet</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:55</td>
<td>lets meet in</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:25:58</td>
<td>i</td>
<td></td>
</tr>
<tr>
<td>22250</td>
<td>1</td>
<td>2011-05-27 08:26:03</td>
<td>so, so wht, so whta, so wh</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:26:04</td>
<td>so</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:26:06</td>
<td>wht</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011-05-27 08:26:07</td>
<td>whta</td>
<td></td>
</tr>
<tr>
<td>Time Stamp</td>
<td>Interaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-05-27 08:25:09</td>
<td>i</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-05-27 08:25:09</td>
<td>in</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-05-27 08:25:10</td>
<td>i need</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-05-27 08:25:20</td>
<td>i need t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-05-27 08:25:21</td>
<td>i need to</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-05-27 08:25:26</td>
<td>i need t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-05-27 08:25:28</td>
<td>i need</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Showing 1 to 7 of 7 entries
## User Activity

**2013/03/01**

- **Number of Queries**: 1
- **Number of Interactions**: 10

### Query Log

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Time Stamp</th>
<th>Interaction(s)</th>
<th>Number of Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>305228</td>
<td>2012-09-05 17:09:37</td>
<td>in the</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:09:37</td>
<td>in the m</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:09:38</td>
<td>in the mid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:09:40</td>
<td>in the middle of s</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:09:40</td>
<td>in the middle of st</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:09:41</td>
<td>in the middle of street</td>
<td></td>
</tr>
<tr>
<td>305229</td>
<td>2012-09-05 17:10:20</td>
<td>isn't n</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:10:21</td>
<td>isn't normal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:10:21</td>
<td>isn't norma</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:10:21</td>
<td>isn't norma</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:10:21</td>
<td>isn't norma</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2012-09-05 17:10:21</td>
<td>isn't norma</td>
<td></td>
</tr>
</tbody>
</table>
Custom Search

Minimum Duration of presence(s):

Minimum Number of queries:

Minimum Number of interactions:

First Appeared: 03/15/2009

Last Appeared: 02/01/2014

search
Ideas

- Learning effect
- Identifiable user
Informative Linguistic Knowledge Extraction from Wikipedia

Roxanne El Baff (1st Semester CSM Student)
Supervisor : Khalid El Khatib
Wikipidea and JWPL

- High quality, up to date knowledge base
- JWPL (Java Wikipedia Library)

Natural Language Processing

- Title
- Content
- Links
Measuring Term Informativeness

Term Informativeness Measurements

Statistic
- Term Frequency
- Document Frequency

Semantic
- Semantic Relatedness

Context-Aware Term Informativeness

Measure the importance of a term based on
- Its context
- Importance of the term (Statistic)
- Importance of its context (How strong is the relation between term and context (Semantic Relatedness )

\[ I(t, c_i) = \sum_{c_j \in U_f(t)} \kappa(c_i, c_j) \cdot CA(c_j) \]
Elasticsearch and the Clueweb

A Work-in-Progress Presentation

Janek Bevendorff

Web Technology & Information Systems
Bauhaus-Universität Weimar
Elasticsearch and the Clueweb

What is the Clueweb?

some data:

- web crawl of 1,040,809,705 documents
- 5TB of compressed data (25TB uncompressed)
- 4,780,950,903 unique URLs
- tons of spam
Elasticsearch and the Clueweb
And what do we do with it?

63481 other results (30.184 seconds) from 32/40 responding nodes

allAfrica.com: Kenya: Country to Launch Obama Family Tree (Page 1 of 1)
http://allafrica.com/stories/200801260024.html

History books, especially on the Luo migration," he said. M.Reria said his department would come up with the draft on Obama's family tree, which would be adopted by the Ministry of Heritage and Culture. In a related development, the Kenyan government sai

Genealogy Ireland - Obama Irish Family Tree
http://www.eneclann.ie/Research/genealogy_Obama_family_tree.html

Genealogy Ireland - Obama Irish Family Tree Michael Kearney, d. 1762 Master of the Guild of Barber Surgeons, 1726 Joseph Kearney, b.ca.1698 buried 20th Jan.

Japan America Society of Greater Philadelphia - Membership
http://jasgp.org/component/page.shop.browse/category_id,1?option.com_virtuemart/Itemid,171/

Read More Previous Blog Entries Obama Manju, Obama Burgers Tokyo No Records [...] $15.00 Product Details Individual/Family Membership $50.00 Product Details [...] - Membership Home About Events Tree Planting Project Membership Blog

Japan America Society of Greater Philadelphia - Student Membership
Blog Entries Obama Manju, Obama [...] Individual/Family Membership [...] About Events Tree Planting

>> more results from jasgp.org
Elasticsearch and the Clueweb

New backend: Elasticsearch

Elasticsearch is a...

- distributed and redundant Lucene index
- (RESTful) search server

Optionally, Elasticsearch comes with Hadoop integration for indexing large amounts of data and performing real-time search on HDFS clusters.
Elasticsearch and the Clueweb
Chatnoir 2

Search results 1-10 for “obama family tree”

clueweb09-en0001-02-21241
...Ancestry of Barack Obama - Family Tree and Genealogy of Senator Obama
clueweb09-en0001-35-11959
..._ARTS SPORTS BUSINESS OPINION CLASSIFIED BLOGS Login Register Contact Subscribe Obituaries E-Edition Photos Videos Fun & Games Gazette Photo Gallery Buy Gazette photos online Christmas trees Several varieties of Christmas trees are available at Elm's Family Tree Farm in Ballston Spa, which is a traditional spot for many holiday tree shoppers. Posted on December 5, 2008. E-mail this gallery to a friend…
clueweb09-en0001-02-21240
... Martin Luther King, Jr. Historic civil rights leader Martin Luther King, Jr. was actually born with the name Michael King, one of the three children born to Martin Luther King, Sr. and Alberta Williams King. Learn about the ancestors and history of Martin Luther King in this only family tree. Barack Obama Learn about the deep African and American roots of Barack Obama, US Senator and presidential candidate. His African roots stretch back for generations in Kenya, while his American roots connect to…
clueweb09-en0001-75-31244
... is my only free night until the weekend and I simply can’t wait that long if I don’t find a tree, the tree, tonight. A little girl’s happy squeals erupt a few trees over. Curious to see which tree has found its family, I amble over. A young girl zipped and hooded inside a pink puffy jacket hops up and down holding her mother’s hand. Her dad gives the tree a final once-over. The little girl hops faster. Her mother tells her gently to calm down. She stands still and pushes the hood off of her…
Future Work

- index the whole ClueWeb12 and ClueWeb09 datasets on our brand new Betaweb cluster
- use more fields (title, URL, anchor texts etc.) for weighted search
- some more frontend magic
Elasticsearch and the Clueweb

Thank you for your attention!
Agenda

1. Passphone Protocol

2. AVISPA
Passphone Protocol

- Protocol for two factor authentication at a service provider
- Factors:
  - Password as usual
  - Smartphone
- User enters his password
- Gets a QR-Code in return
- Scans the QR-Code with his registered smartphone app
- After success the user is logged in
Passphone Protocol

- Protocol for two factor authentication at a service provider
- Factors:
  - Password as usual
  - Smartphone
- User enters his password
- Gets a QR-Code in return
- Scans the QR-Code with his registered smartphone app
- After success the user is logged in
In protocol: several communications with different parties:
- Service Provider (e.g. Facebook, Ebay, Amazon, ...)
- Trusted Third Party Server
- User at a browser
- User at his smartphone

Communication save?
Approach: automatic proofing of the protocol with AVISPA

AVISPA = Automated Validation of Internet Security Protocols and Applications

Protocol has to be translated into special language HLPSL

HLPSL = High Level Protocol Specification Language
- Approach: automatic proofing of the protocol with AVISPA
- AVISPA = Automated Validation of Internet Security Protocols and Applications
- Protocol has to be translated into special language HLPSL
- HLPSL = High Level Protocol Specification Language
AVISPA Function

High-Level Protocol Specification Language (HLPSL)

Translator
HLPSL2IF

Intermediate Format (IF)

On-the-fly Model-Checker
OFMC

CL-based
Attack Searcher
CL-AtSe

SAT-based
Model-Checker
SATMC

Tree Automata-based
Protocol Analyser
TA4SP

Output
Possible to choose the proofer
output after proof depends on criteria set in the hlspl file
(e.g. security of a nonce)
Proofer checks if the given protocol is safe or if not
If a Protocol is not safe the proofer gives an attack trace
135 Server

= 27 x
Disk Space

2160 TB
10GbE Network
1620 Cores

8.64 TB RAM
Network Boot

Scanning for devices. Please wait, this may take several minutes...

Intel(R) Boot Agent XE v2.3.08
Copyright (C) 1997-2013, Intel Corporation

CLIENT MAC ADDR: EC F4 BB C9 35 B2 GUID: 44454C4C 5400 104D 8030 C4C04F573232
CLIENT IP: 141.54.132.1 MASK: 255.255.255.0 DHCP IP: 141.54.65.1
GATEWAY IP: 141.54.132.254
XPXE entry point found (we hope) at 9837:0106 via plan A
UNDI code segment at 9837 len 4810
UNDI data segment at 90F5 len 7420
Getting cached packet 01 02 03
My IP address seems to be 8D368401 141.54.132.1
ip=141.54.132.1:141.54.132.20:141.54.132.254:255.255.255.0
BOOTIF=01-ec-f4-bb-c9-35-b2
SYSSUID=44454c4c-5400-104d-8030-c4c04f573232
TFTP prefix: /tftpboot/
Trying to load: pxelinux.cfg/default ok
BETAWEB
boot:
Loading vmlinuz........
Loading initrd.lz.........................ready.
Setup via Configuration Management
SimHash as a Service

Scaling Near-Duplicate Detection

Jan Graßegger
Freiwilligentag in Weimar: Ein Tag mit 680 Stunden

23.09.2014 - 11:00 Uhr

Wenn 170 Leute vier Stunden lang ranklotzen, schaffen sie viel mehr als ein Arbeiter in 680 Stunden. Gemeinsamkeit macht stärker, deshalb sind die Einsätze an den Weimarer Freiwilligentagen nicht mit Geld aufzuwerten.

ZUM THEMA
Rekord an guten Taten beim Freiwilligentag in Jena
Rekordbeteiligung beim 10. Freiwilligentag: Mehr als 320 Freiwillige kürmerten sich an 34 Ein... mehr

Freiwilligentag in Weimar: Ein Tag mit 680 Stunden

Erfurter Freiwilligentag für das Gemeinwohl
Kindertag mit Spaß und Wünschen in Gera
Erster Freiwilligentag in Eisenach
Bilder des Tages im Monat September
Erster Thüringer Freiwilligentag im Landkreis Nordhausen
Thüringer Freiwilligentag im Landkreis Nordhausen
Im Nachbarschaftszentrum in Eisenach wird beraten und auch gemeinsam gekocht
Helfer für Freiwilligentag am 26. September in Gera gesucht

23.09.2014 - 11:00 Uhr
Wenn 170 Leute vier Stunden lang ranklotzen, schaffen sie viel mehr als ein Arbeiter in 680 Stunden. Gemeinsamkeit macht stärker, deshalb sind die Einsätze an den Weimarer Freiwilligentagen nicht mit Geld aufzuwerten.

Bäumchen, nützlich: Auf der Streuselwiese bei Gabendorf hängt kein Apfel mehr am Baum, nachdem die Grüne Liga mit zehn Helfern alles erntete, was sich zu Saft verarbeiten lässt. Foto: Sabine Brandt

Weimar. "Wir haben alles geschafft, was wir uns vorgenommen hatten", zieht Stefanie Lachmann von der Ehrenamtsagentur zufrieden Bilanz unter Weimars größten Subbotnik, der am zurückliegenden Samstag ausgerufen worden war. Seither sind der Lebenshilfleider um einen Bildersieder und die Grüne Liga um zwei Tonnen Obst aus Gabendorf für die Saftpresse reich und das Schlachthofviertel um einige Kilo Müll.
SimHash \[\text{[Cha02]}\]

- Locality-Sensitive Hash
- embeds document text into a 64-bit hash
- correlates with Cos-Similarity
SimHash as a Service

Searching for near-duplicates over a web service

- corpus: ClueWeb12 (over 700M docs)
- response time: < 1 second
- search tables allow fast candidate retrieval [MJS07]
- works with aitools-invertedindex3

One class classification of vandalism in the wikipedia

Speaker:
Jonas Köhler
The classification problem:

Classify edits of wikipedia entries into regular edits and vandalism edits.

- Currently he is the Chairman of the [[World of Labor Institute]].

+ Currently he is the Chairman of the [[World of Labor Institute]], and wants to breed an army of termites to claim world domination.

The corpora:

PAN WVC 2010 and PAN WVC 2011¹ (humanly annotated edits: vandalism and regular)

PAN WVC 2010

2394 vandalism entries ⇒ imbalanced classes...
30045 regular entries

Features: 54 few meta-data, few linguistic data ⇒ dimensionality will grow!

One class classification

Train a model with data of the **positive class only**.

The model shall detect if a data vector is **positive or an outlier** from this class.

Useful if: the negative class is hard to describe with feature model

the negative class is difficult to sample

the class cardinality is very imbalanced

⇒ There are two ways Wikipedia vandalism detection can be seen as a OCC:

1) vandalism can be modelled with features
   regular entries probably can’t  \[\text{positive} = \text{vandalism}\]

2) a lot more regular entries exist
   annotation of vandalism entries is expensive  \[\text{positive} = \text{regular}\]
Outlook

**What we have tried:**

applying two standard implementations (libsvm)

applying a method intended for high dimension OCC (based on Random Forest \(^1\))

**Results**

standard implementations do not work on PAN-WVC-2010 and PAN-WVC-2011

there is a lot of research on OCC, but only few implementations of methods are available

implementing the methods by our own is not feasible

**How we want to proceed now:**

continue with the work on the features (meta-data, NLP, ...)

analyze the „hard cases“ (there are ~280 entries which are always bad in recall)

\(^1\) Chesner Désir, Simon Bernard, Caroline Petitjean, Heutte Laurent. One class random forests. Pattern Recognition, Elsevier, 2013, 46, pp.3490-3506.
Thank you!

Questions?