

Interactive Text Exploration

Günter Neumann,
DFKI, Saarbrücken, Germany
Joined work with Sven Schmeier, DFKI, Berlin.

+ Overview of my talk

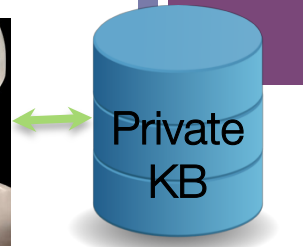
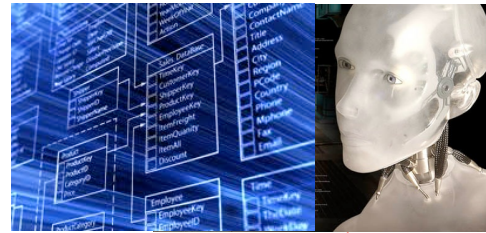


- Motivation and Background
- Interactive exploratory search
- Methods and technology
- Where we are, where we want to go

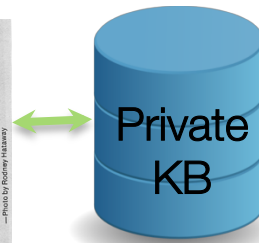
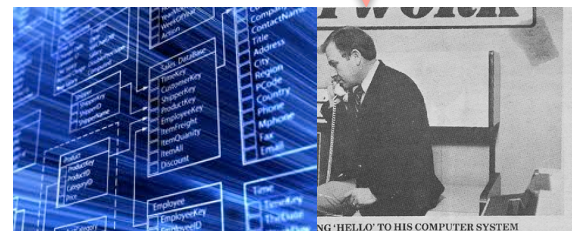
+ “The Big Idea”

- The **extraction**, **classification**, and **talking about information** from large-scale unstructured noisy multi-lingual text sources.

Topic of Interest



Text as interface

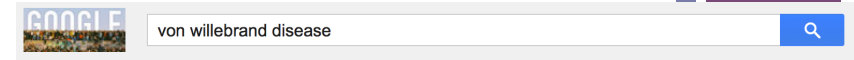


„Reading text and talking about it“



Motivation

- Today's Web search is still dominated by one-shot-search:
 - Users basically have to know what they are looking for.
 - The documents serve as answers to user queries.
 - Each document in the ranked list is considered independently.
- Restricted assistance in content-oriented interaction



Web Images Videos Books News More Search tools

About 783,000 results (0.27 seconds)

[Von Willebrand disease - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Von_Willebrand_disease

Von Willebrand disease (vWD) is the most common hereditary coagulation abnormality described in humans, although it can also be acquired as a result of other ...

[Signs and symptoms](#) - [Diagnosis](#) - [Classification and types](#)

[What Is von Willebrand Disease? - NHLBI, NIH](#)

www.nlm.nih.gov/healthinformationforthepublic/healthtopics/

Von Willebrand disease (VWD) is a bleeding disorder. It affects your blood's ability to clot. If your blood doesn't clot, you can have heavy, hard-to-stop bleeding ...

[Von Willebrand disease - Mayo Clinic](#)

www.mayoclinic.org/diseases-conditions/von-willebrand-disease/overview/mon-20030195

Von Willebrand disease is a condition that can cause extended or excessive bleeding.

The condition is most often inherited but in rare cases may develop later ...

[OMIM Entry - # 193400 - VON WILLEBRAND DISEASE ...](#)

www.omim.org/entry/193400

The classification of **von Willebrand disease** has a long and complex history. The current classification is based on that described by Sadler (1994) and updated ...

Searches related to von willebrand disease

[von willebrand factor](#)

[von willebrand disease treatment](#)

[von willebrand disease diagnosis](#)

[von willebrand disease genetics](#)

[von willebrand disease symptoms](#)

[von willebrand disease and pregnancy](#)

[von willebrand disease emedicine](#)

[von willebrand disease in dogs](#)



1 2 3 4 5 6 7 8 9 10

Next

+ Exploratory Search



- We consider a **user query** as a specification of a **topic** that the user wants to know and learn more about. Hence, the **search result** is basically a **graphical structure of the topic and associated topics** that are found.
- The **user can interactively explore this topic graph** using a simple and intuitive (touchable) user interface in order to either learn more about the content of a topic or to interactively expand a topic with newly computed related topics.

+ Exploratory Search on Mobile Devices

14:36 35%
i-GNSSMM+

jim clark GO!

14:36 34%
i-GNSSMM+

Search ON THIS DAY by date 7 April GO

ON THIS DAY 1950 2005 **7 April** BBC NEWS

Search ON THIS DAY by date 7 April GO

Front Page | Years | Themes | Witness

About This Site | Text Only

1968: Jim Clark killed in car smash

Motor racing world champion Jim Clark has been killed in a car crash during a Formula Two race at Hockenheim.

Clark, 32, was at the wheel of his Lotus-Cosworth which left the track at 170mph (274km/h), somersaulted through the air and collided with a tree on a remote part of the German track.

The twice Formula One champion, who sustained a broken neck and a fractured skull, was dead before he reached hospital.

The cause of the accident is not yet known although experts have suggested it could have been a fault in the steering mechanism or rear-axle suspension.

Although it had been raining prior to the race, this is not thought to have caused Clark's car to skid.

"The car seemed to be in a thousand pieces"
Eye-witness

The 80,000 spectators, who were informed of the accident via loudspeaker some two hours later, were stunned by the news.

They spontaneously rose to their feet in silent tribute.

The only witness to the accident was a track marshal who said: "I was horror-struck. Everything

Jim Clark was regarded by many as one of the safest drivers in the sport

In Context

Jim Clark was one of more than 100 international racing drivers killed 'in action' between 1958 and 1968.

Clark, who is still considered by some as the most natural racing driver of all time, won the World Championship twice - first in 1963 and then in 1965.

He won 25 Grand Prix races and was the first Briton to win the gruelling Indianapolis race in America.

He was made an OBE for his services to motor racing in 1964.

Stories From 7 Apr

- 1978: Carter delays N-bomb production
- 1976: Government crisis as Stonehouse quits
- 1968: Jim Clark killed in car smash
- 1986: Sinclair sells computer business
- 1999: US claims 'banana war' victory

formula one hero
formula one driver
texas
formula one driver
texas hideout
texas hideout

i-GNSSMM i-GNSSMM+ i-GNSSMM+

o2-de 3G 15:40
i-GNSSMM+

jim clark GO

o2-de 3G 15:41
i-GNSSMM+ jim clark

driver (4)

texas hideout (1)

16:16

Snippets

Q W E R T Z
A S D F G H J K L
Y X
123

o2-de 3G 16:16

JIMMY CLARK was born in Kilmany, in the county of Fife to a Scottish farming family, roots that would stay with him for the rest of his life. He was the only son in a family of four daughters. His early racing exploits were initially met by family disapproval, not the sort of activity suitable for someone expected to take over the family farm. Clark, undeterred continued to race in rallies and other local races under the guidance of his close friend Ian Scott-Watson. Later he joined a team run by Jock McBain known as the Border Reivers named after raiders who plagued the border regions between Scotland and England during the 13th-16th centuries. According to fellow Border Reiver Ian Scott Watson, "Jim drove so fast that most people were scared stiff to sit next to him." In one of these races Clark drove a Lotus Elite against none other than [Colin Chapman](#), who was a fair driver in his own right. Chapman was very impressed by the young Scotsman and he would keep an eye on this young lad. Ironically in 1959 the Border Reivers planned to buy a single-seater Formula 2 Lotus for Clark but after watching [Graham Hill](#) lose a wheel in a similar car, Clark decided that the Lotus cars were not safe and that he would stick to sports cars for the time being. Eventually he graduated to an Aston Martin which brought him to the attention of Reg Parnell, the factory team

Allstate Auto, Life and Home Insurance. You're in Good Hands. Call NOW to Speak With an Agent



Our Approach – On-demand Interactive Open Information Extraction



- Topic-driven Text Exploration
 - Search engines as API to text fragment extraction (snippets)
- Dynamic construction of topic graphs
 - Empirical distance-aware phrase collocation
 - Open relation extraction
- Interaction with topic graphs
 - Inspection of node content (snippets and documents)
 - Query expansion and eventually additional search
 - Guided exploratory search for handling topic ambiguity



Search

von Willebrand

The von Willebrand

Type 2 von Willebrand

Von Willebrand

Intracellular structure

Acquired von Willebrand

Porcine and canine studies.

Pregnancy and

Investigation of

Multiple von Willebrand

Oligosaccharic

o2 - de 10:33 34% i-GNSSM++

von willebrand disease

von willebrand disease

GO!

DFKZ

von willebrand disease type I

von willebrand's disease type

von willebrand's disease

inherited

bleeding disorder caused by decrease

is, is, account

is

von willebrand disease

is characterized

platelet glycoprotein

missense

coincides

von willebrand factor gene

von willebrand factor pseudogene

von willebrand factor

von willebrand factor locus

von willebrand disease

The image shows a mobile application interface for searching and visualizing relationships between terms. At the top, there are search input fields containing 'von willebrand disease' and a 'GO!' button. Below the search bar is a network diagram with 'von willebrand disease' as a central node. It is connected to several other nodes: 'bleeding disorder caused by decrease', 'is, is, account', 'is', 'von willebrand disease' (a self-loop), 'is characterized', 'platelet glycoprotein', 'missense', 'coincides', 'von willebrand factor gene', and 'von willebrand factor pseudogene'. On the left side, there is a vertical list of search results or related terms, including 'von Willebrand', 'The von Willebrand', 'Type 2 von Willebrand', 'Von Willebrand', 'Intracellular structure', 'Acquired von Willebrand', 'Porcine and canine studies.', 'Pregnancy and', 'Investigation of', 'Multiple von Willebrand', and 'Oligosaccharic'. The bottom of the screen shows a standard mobile OS navigation bar with icons for home, search, and other functions.

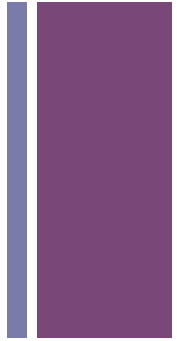
+ Topic Graphs



- Main data structure
 - A graphical summary of relevant text fragments in form of a graph
 - Nodes and edges are text fragments
 - Nodes: entities phrases
 - Edges: relation phrases
 - Content of a node: set of snippets it has been extracted from, and the documents retrievable via the snippets' web links.
- Properties
 - Open domain
 - Dynamic index structure
 - Weight-based filtering/construction



Construction of Topic graphs



- Identification of relevant text fragments
 - A document consisting of topic-query related text fragments
- Identification of nodes and edges
 - Distance-aware collocation
 - Clustering-based labels for filtering
- Technology
 - **Shallow** Open relation Extraction (ORE) for snippets
 - **Deeper** ORE for more regular text

Chunk-pair distance model

Topic pair weighting

Topic graph visualization

For each chunk c_i do:

$(c_i, c_{i+1}, d_{i(i+1)}), (c_i, c_{i+2}, d_{i(i+2)}), \dots$

$(c_i, c_j, \#c_i, \#c_j, D_{ij})$ with

$D_{ij} = \{(freq_1, dist_1), (freq_2, dist_2), \dots\}$

$$PMI(cpd) = \log_2\left(\frac{p(c_i, c_j)}{p(c_i) * p(c_j)}\right)$$

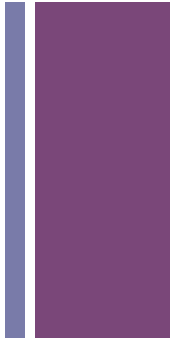
$$= \log_2(p(c_i, c_j)) - \log_2(p(c_i) * p(c_j))$$



The visualized topic graph TG is then computed from a subset $CPD'_M \subset CPD_M$ using the m highest ranked cpd for fixed c_i . In other words, we restrict the complexity of a TG by restricting the number of edges connected to a node.



Evaluation of Mobile Touchable User Interface



- 20 testers
 - 7 from our lab
 - 13 “normal” people
- 10 topic queries
 - Definitions: EEUU, NLF
 - Person names: Bieber, David Beckham, Pete Best, Clark Kent, Wendy Carlos
 - General: Brisbane, Balancity, Adidas.
- Average answer time for a query: ~0.5 seconds

Table 1: Google

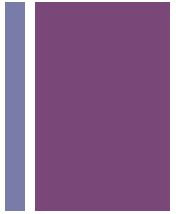
#Question	v.good	good	avg.	poor
results first sight	55%	40%	15%	-
query answered	71%	29%	-	-
interesting facts	33%	33%	33%	-
suprising facts	33%	-	-	66%
overall feeling	33%	50%	17%	4%

Table 2: i-GNSSMM

#Question	v.good	good	avg.	poor
results first sight	43%	38%	20%	-
query answered	65%	20%	15%	-
interesting facts	62%	24%	10%	4%
suprising facts	66%	15%	13%	6%
overall feeling	54%	28%	14%	4%



Guided Exploratory Search



- Problem: a topic graph might merge information from different topics/concepts
- Solution:
 - Guided exploratory search
 - Using an external KB (e.g., Wikipedia)
- Strategy
 - Compute topic graph TD_q for query q
 - Ask KB (Wikipedia or any other KB) if q is ambiguous
 - Let user select reading r , and use selected Wikipedia article for expanding q to q'
 - Compute new topic graph $TD_{q'}$





Clark was associated with various criminalRobert Big Bob Brady and

14:45 Lädt nicht

i-GNSSMM

GO!

```

    graph TD
      JR(jochen rindt) --- IC(im clark)
      JR --- GH(graham hill)
      JR --- MH(mike hawthorn)
      JR --- R(race)
      MS(mike spence) --- IC
      MS --- GH
      MS --- MH
      MS --- R
      GH --- IC
      GH --- MH
      GH --- R
      IC --- MH
      IC --- R
      MH --- R
  
```

Home i-GNSSMM i-GNSSMM+ i-GNSSMM++ i-GNSSMM i-GNSSMM+ i-GNSSMM++ i-Wiki i-WebQA Settings About

+ Evaluation



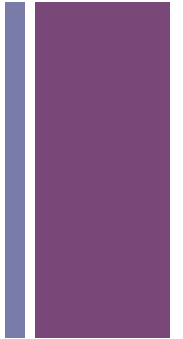
List of celebrity guest stars in Sesame Street:

209 different queries



List of film and television directors:

229 different queries





Evaluation



- Goal:
 - We want to analyze whether our approach helps building topic graphs which **express a preference** for the selected reading.

- Automatic evaluation:
 - Method
 - For each reading article r , compute topic graph TD_r using expanded query
 - Compare TD_r with all readings and check whether best reading equals r
 - Advantage: No manual checking necessary
 - Disadvantage: Correctness of TD_R needs to be proven

- Manual evaluation:
 - Double-check the results of the automatic evaluation
 - Prove the results at least for the examples used in evaluation

+ Results

set	#queries	good	bad	acc
Sesame + Colloc.	209	375	54	87.41 %
Sesame + Colloc.+ SemLabel	209	378	51	88.11 %
Hollywood + Colloc.+ SemLabel	229	472	28	94.40 %
Hollywood + Colloc.+ SemLabel	229	481	19	96.20 %

1 st task		2 nd task			
set	guidance	associated topics	good	bad	accuracy
Sesame	ca. 95 %	167	132	35	79.04 %
Hollywood	ca. 95 %	145	129	16	89.00 %
Sesame	> 97 %	167	108	59	64.67 %
Hollywood	> 97 %	145	105	40	72.41 %

Automatic

- Colloc. – empirical collocations for topic graph computation
- SemLabel – Filtering of nodes using semantic labels computed via SVD (Carrot2)

Manual

- 2 test persons
- 20 randomly chosen celebrities and 20 randomly chosen directors
- 1st task: Exploratory search and personal judgments of the Guidance by the system
- 2nd task: Check all associated nodes after choosing a meaning in the list



+ Summary and Discussion



- Interactive topic graph exploration
 - Unsupervised open information extraction
 - On-demand computation of topic graphs
 - Strategies for guided exploratory search
 - Effective for Web snippet like text fragments
 - Implemented for EN and DE on mobile touchable device
- Drawback
 - Problems in processing text fragments from large-scale text directly
 - Especially Open Relation Extraction for German is challenging
- Solution:
 - Nemex - A new multilingual Open Relation Extraction approach





Nemex – A Multilingual Open Relation Extraction Approach



- Uniform multilingual core ORE
 - N-ary extraction
 - Clause-level
- Multi-lingual
 - Very few language-specific constraints over dependency trees
 - Current: English and German
- Efficiency
 - Complete pipeline (from sentence splitting, to POS-tagging, to NER, to dependency parsing, to relation extraction)
 - About 800 sentences/sec
 - Streaming based – small memory footprint



German ORE is Challenging

YES WE CAN!



- Challenging properties of German
 - Morphology/Compounding*
 - No strict word ordering (especially between phrases)
 - Discontinuous elements, e.g., verb groups
- Simple, pattern-based ORE approach difficult to realize (Chen&Manning, 2014)
- Deep sentence analysis helpful
 - Current multilingual dependency parsers provide very good robustness!
 - DFKI's MDParser is very efficient: 1000sentences/second (but see also Chen&Manning, 2014)
- Challenge:
 - Can we design a core uniform ORE approach for English, German, ... ?

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz
"the law concerning the delegation of duties for the supervision
of cattle marking and the labelling of beef"



Multilingual ORE – Our Approach



- Multi-lingual open relation extraction
 - Only few Language-specific constraints necessary (constraints over direct dependency relations (head, label, modifier))
 - Few language-independent constraints in case of uniform dependency annotations, e.g., McDonald et al., 2013
- Processing strategy
 - Head-Driven Phrase Extraction
 - Top-down head-driven traversal of dependency tree

+ Example: English



Mammalian NMD was mostly studied in cultured cells so far and there was no direct evidence yet that NMD could operate in the brain .

Dependency
Tree (uniform tag
and label set;
Conll format):

1:Mammalian:NOUN:compmod:2
2:NMD:NOUN:nsubjpass:5
3:was:VERB:auxpass:5
4:mostly:ADV:advmod:5
5:studied:VERB:ROOT:0
6:in:ADP:adpmod:5
7:cultured:ADJ:amod:8
8:cells:NOUN:adpobj:6
9:so:ADV:advmod:10
10:far:ADV:advmod:5
11:and:CONJ:cc:5
12:there:DET:expl:13
13:was:VERB:conj:5
14:no:DET:det:16
15:direct:ADJ:amod:16
16:evidence:NOUN:nsubj:13
17:yet:ADV:advmod:13
18:that:ADP:mark:21
19:NMD:NOUN:nsubj:21
20:could:VERB:aux:21
21:operate:VERB:advcl:13
22:in:ADP:adpmod:21
23:the:DET:det:24
24:brain:NOUN:adpobj:22
25:::p:5

+ Example English – cont.

*

(Mammalian NMD, was mostly studied so far, in cultured cells)
(no direct evidence, was yet, there)
(NMD, could operate, in the brain)

**Annotated sentence:

[[[Arg11 Mammalian NMD Arg11]]] --->Rel1 was mostly studied
[[[Arg13 in cultured cells Arg13]]] so far Rel1<--- and [[[Arg23 there
Arg23]]] --->Rel2 was [[[Arg21 no direct evidence Arg21]]] yet
Rel2<--- that [[[Arg31 NMD Arg31]]] --->Rel3 could operate Rel3<---
[[[Arg33 in the brain Arg33]]] .

*Details omitted

**Extension of the annotation scheme introduced by Mesquita et al., 2013

+ Example: German

Zuvor hatte Asmussen mitgeteilt, dass er sein Amt als EZB-Direktor in Kürze aufgeben will:

**Earlier had Asmussen informed, that he his position as EZB-director in the_near_future quit will:*

Earlier Asmussen has informed that he will quit his position as EZB-director in the_near_future:

Dependency
Tree (uniform tag
and label set;
Conll format):

1:Zuvor:ADV:advmod:2
2:**hatte**:VERB:ROOT:0
3:Asmussen:NOUN:nsubj:2
4:**mitgeteilt**:VERB:aux:2
5:,::p:2
6:dass:CONJ:mark:14
7:**er**:PRON:nsubj:14
8:sein:PRON:poss:9
9:Amt:NOUN:dobj:14
10:als:ADP:adpmod:14
11:EZB-Direktor:NOUN:adpobj:10
12:in:ADP:adpmod:14
13:Kürze:NOUN:adpobj:12
14:**aufgeben**:VERB:NMOD:2
15:**will**:VERB:aux:14
16:::..NMOD:2

+ Example German – Cont.

(Asmussen, Zuvor hatte mitgeteilt)

(er, aufgeben will, sein Amt, als EZB-Direktor, in Kürze)

Annotation:

--->Rel1 Zuvor hatte [[[Arg1 1 Asmussen Arg1 1]]] mitgeteilt Rel1<--- ,
dass [[[Arg21 er Arg21]]] [[[Arg22 sein Amt Arg22]]] [[[Arg23 als EZB-
Direktor Arg23]]] [[[Arg24 in Kürze Arg24]]] --->Rel2 aufgeben will
Rel2<--- :

+ Nemex – Current Status



- Properties
 - Efficient text stream for EN and DE implemented
 - Uniform POS and Dependency labels
 - Small set of uniform constraints over dependency relations
- Very fast & Domain independent
 - About 800 sentences per second for complete pipeline
- Current /near future work
 - Improve cross-clausal resolution
 - Extensive evaluation, intrinsic and extrinsic
 - Adaptation to other languages
 - Conll based dependency treebanks (uniform and specific)

+ Future action points



- Cross-sentence open information extraction
 - **Goal: co-reference resolution, integration of more fine-grained information to dependency parsers (morphology), text inference**
- Beyond isolated topic graphs
 - **Goal: share topic graphs, compare topic graphs, monitor topic graphs**
- Interactive text data mining and knowledge discovery
 - **Goal: support abstract interactions, e.g., “more like this”, “less like this”, “what is this”, ...**



DONE

Thank you for Your Attention !