

Information Retrieval

PS Einführung in die Computerlinguistik
SE aus Artificial Intelligence

Dieter Merkl

dieter.merk1@ec.tuwien.ac.at

Electronic Commerce Group

Institute of Software Technology and Interactive Systems
Vienna University of Technology
Favoritenstraße 9-11/188-1 . 1040 Vienna . Austria/Europe
Fax: +43 (1) 58801 - 18899
<http://www.ec.tuwien.ac.at/~dieter/>



Um was geht's da jetzt eigentlich?

- Ganz pragmatisch ... es geht ums Auffinden von Texten
- ... kann sein im Sinne von "ich suche was und hätte gerne Hinweise auf Quellen, wo vielleicht was darüber drinnen steht"
-> kommt das bekannt vor?
- ... kann auch sein im Sinne von "ich hab schon etwas, das ganz hilfreich ist, hätte jetzt aber gern mehr dazu"
- ... kann aber auch sein im Sinne von "ich würde jetzt doch ganz gerne wissen, wie dieses Thema in Bezug zu anderen steht"

Basic approach to IR (*)

- Most successful approaches are statistical
 - Directly, or an effort to capture and use probabilities
- What about natural language understanding?
 - i.e. computer "understands" documents and queries
 - difficult in unrestricted domains
 - can be successful in predictable settings
- What about manually assigned headings?
 - e.g. Dewey Decimal Classification
 - human agreement is not good
 - hard to predict which headings are "interesting"
 - expensive

(*) Tut mir jetzt echt leid, aber ab und zu wird die eine oder andere Folie in Englisch sein - oder vielleicht wird's auch eher so sein, das ab und zu mal eine Folie in Deutsch vorbeikommt :-)

Relevant items are similar

- Much of information retrieval depends upon the idea that
similar vocabulary => relevant to same queries
- or more general
similar vocabularies => similar documents

“Bag of Words”

- An effective and popular approach
- Compares words without regard to order
- Consider reordering words in a headline
 - Random: beating takes points falling another Dow 355
 - Alphabetical: 355 another beating Dow falling points takes
 - “Interesting”: Dow points beating falling 355 takes another
- Actual: Dow takes another beating, falling 355 points

Guess what's this about?

- 16 x said, 14 x McDonalds, 12 x fat, 11 x fries,
- 8 x new, 6 x company french nutrition,
- 5 x food oil percent reduce taste Tuesday,
- 4 x amount change health Henstenburg make obesity,
- 3 x acids consumer fatty polyunsaturated US,
- 2 x amounts artery Beemer cholesterol clogging director down eat estimates expert fast formula impact initiative moderate plans restaurant saturated trans win,
- 1 x added addition adults advocate affect afternoon age Americans Asia battling beef bet brand Britt Brook Browns calorie center chain chemically ... crispy customers cut ... vegetable weapon weeks Wendys Wootan worldwide years York

The (start of the) original text

■ McDonald's slims down spuds

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no.

"It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

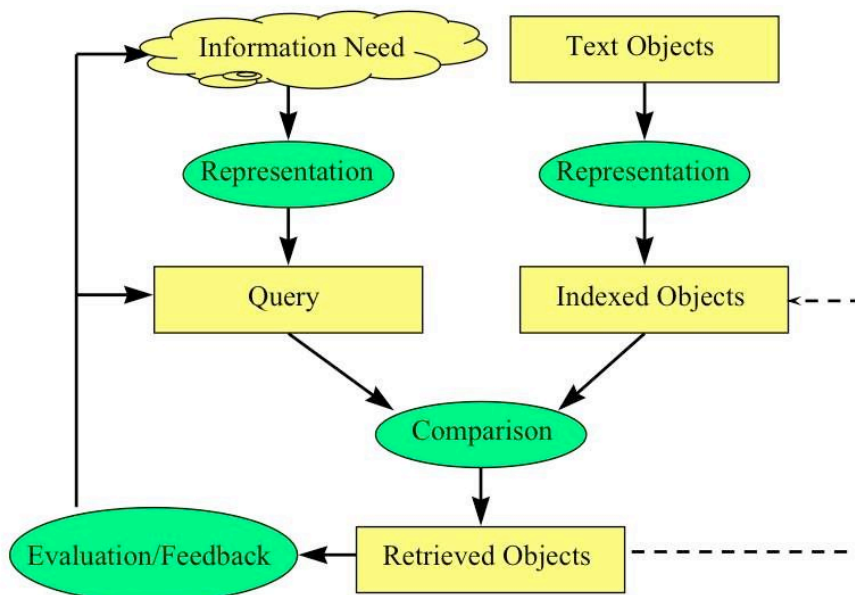
But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down \$0.54 to \$23.22, Research, Estimates) were lower Tuesday afternoon.

...

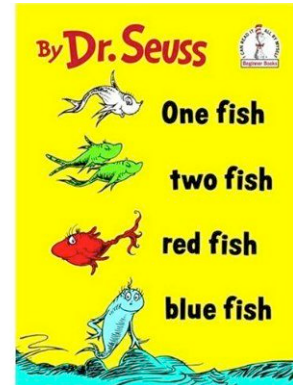
[<http://money.cnn.com/2002/09/03/news/companies/mcdonalds/index.htm>]

Generic view on IR



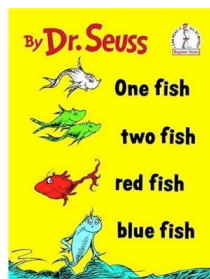
Example: Small document

- $D = \{\text{one fish, two fish, red fish, blue fish, black fish, blue fish, old fish, new fish}\}$
- $\text{len}(D) = 16$
- $P(\text{fish}|D) = 8/16 = 0.5$
- $P(\text{blue}|D) = 2/16 = 0.125$
- $P(\text{one}|D) = 1/16 = 0.0625$
- ...
- $P(\text{eggs}|D) = 0/16 = 0$



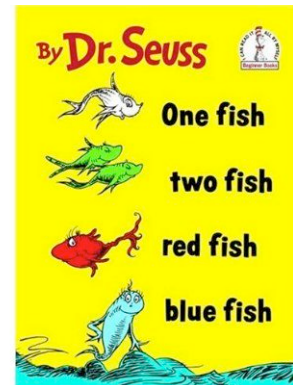
Example: Three small documents

- $D_1 = \{\text{This one, I think, is called a Yink. He likes to wink, he likes to drink.}\}$
- $D_2 = \{\text{He likes to drink, and drink, and drink. The thing he likes to drink is ink.}\}$
- $D_3 = \{\text{The ink he likes to drink is pink. He likes to wink and drink pink ink.}\}$
- Query "drink"
 - $P(\text{drink}|D_1) = 1/16$
 - $P(\text{drink}|D_2) = 4/16$
 - $P(\text{drink}|D_3) = 2/16$
- Query "pink ink"
 - $P(\text{pink ink}|D_1) = 0 \cdot 0 = 0$
 - $P(\text{pink ink}|D_2) = 0 \cdot 1/16 = 0$
 - $P(\text{pink ink}|D_3) = 2/16 \cdot 2/16 \approx 0.016$
- Query "wink drink"
 - $P(\text{wink drink}|D_1) = 1/16 \cdot 1/16 \approx 0.004$
 - $P(\text{wink drink}|D_2) = 0$
 - $P(\text{wink drink}|D_3) = 1/16 \cdot 2/16 \approx 0.008$



Danke für den Hinweis während des Vortrags!

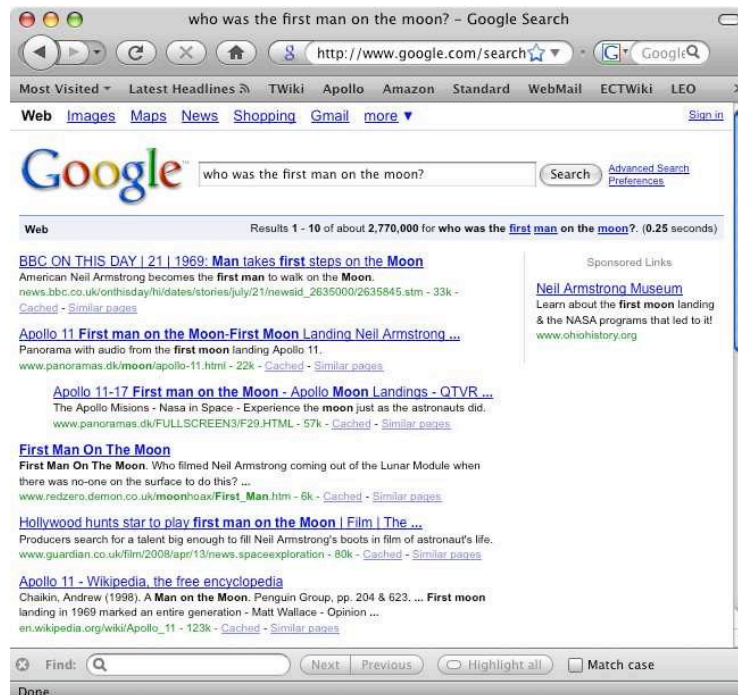
- Die Stelle ist wohl wirklich aus "One fish, two fish, red fish, blue fish" ...
... und nicht wie fälschlicherweise behauptet aus "Green eggs and ham"
 - This one, I think, is called a Yink.
 - He likes to wink,
 - he likes to drink.
 - He likes to drink, and drink, and drink.
 - The thing he likes to drink is ink.
 - The ink he likes to drink is pink.
 - He likes to wink and drink pink ink.
 - SO...
 - if you have a lot of ink,
 - then you should get
 - a Yink, I think.



Basic automatic indexing

- Parse documents to recognize structure
 - e.g. title, date, author, etc
- Scan for word tokens
 - numbers, special characters, hyphenation, capitalization, etc
 - languages like Chinese need segmentation
 - record positional information for proximity operations
- Stopword removal
 - based on short list of common words
 - e.g. articles, conjunctions (the, and, or, ...)
 - saves storage overhead of very long indexes
 - can be dangerous
 - e.g. "to be or not to be", "the who"

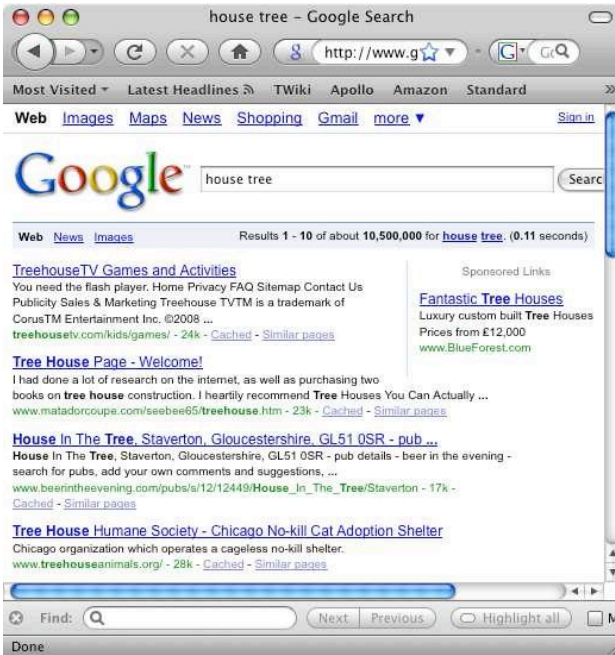
Who was the first man on the moon?



Basic automatic indexing

- Stem words
 - morphological processing to group word variants
 - e.g. plural, declinations
 - can make mistakes but generally preferred
 - not done (or done very carefully) by most Web search engines
- Weight words
 - want more "important" words to have higher weight
 - using frequency in documents and database
 - frequency data independent of retrieval model
- Optional
 - phrase indexing
 - thesaurus classes
 - ...

house tree vs houses trees



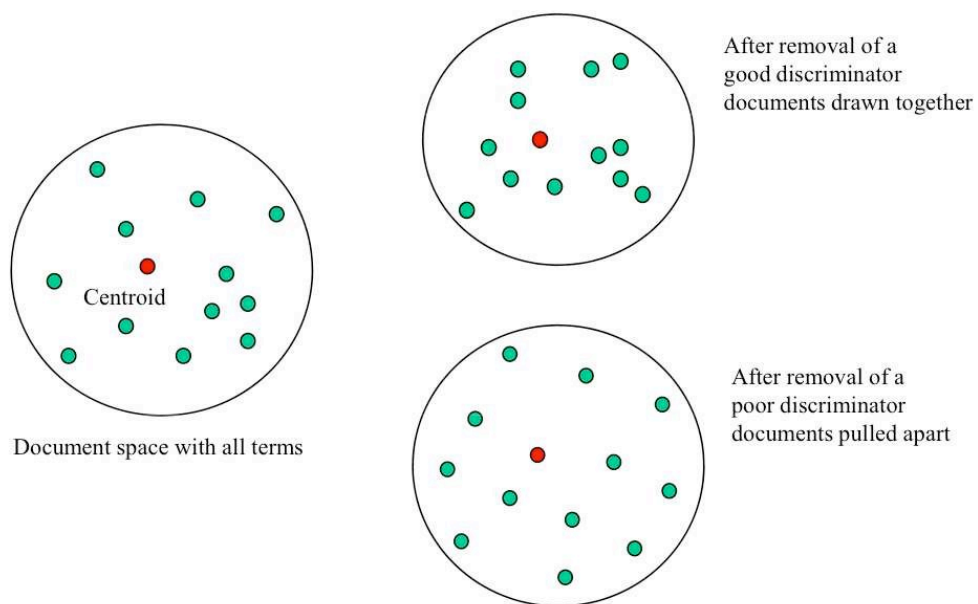
Indexing models

- What makes a term good for indexing?
- What makes an index term good for a query?

Term discrimination model

- Proposed by Gerard Salton in 1975
- Based on vector space model
 - documents and queries are vectors in an n -dimensional space for n index terms
- Compute discrimination value of an index term
 - degree to which the use of the term will help to distinguish documents
- Compare average similarity of documents both with and without an index term

Term discrimination model



Some discriminators for 3 collections

Cranfield 424 (aerodynamics)	MED 450 (medical)	Time 425 (news from 1963)
Best Discriminators		
panel flutter jet cone separate shell yaw nozzle transit degree	marrow Amyloidosis Lymphostasis Hepatitis Hela antigan chromosome irradiate tumor virus	Buddhist Diem Lao Arab Viet Kurd Wilson Baath Park Nenni
Worst Discriminators		
equate theo bound effect solution method press result number flow	clinic children act high develop treat increase result cell patient	work lead Red minister nation party commune U.S. govern new

Term frequency (TF)

- Intuition - the more often a term occurs in a document, the more important it is in describing that document
- Notation: tf_{ij} , i.e. occurrence frequency of term i in document j
- $w_{ij} = tf_{ij}$
- Pro
 - still simple to realize
- Con
 - "length" of document is not taken into account
 $tf_{ij} = 15$ obviously has a different quality in a document containing 100 words or a document containing 10,000 words

Normalized term frequency

- We're getting closer :-)
- Normalization factor for term frequency is used
 - e.g. document length (sum of tf_{ij}), or based on maximum term frequency
 - logarithms used to smooth numbers for large collections
- Most simple form

$$w_{ij} = \frac{tf_{ij}}{\sum_{k=1}^n tf_{kj}}$$

- Con
 - term distribution statistics for the whole document collection is not taken into account
 - e.g. a term appearing frequently in every document is probably less important than a term appearing only in a small number of documents

Inverse document frequency (IDF)

- IDF - inverse document frequency
- Normalization factor for the characteristics of term distribution in the whole document collection
- Intuition
 - good index terms appear frequently within the document, yet rarely within the collection
 - index terms that appear in many documents of the collection are not overly helpful when trying to discriminate between documents (c.f. term discrimination model)

TF·IDF

- We're there, at last :-)
- Notation
 - df_i , i.e. document frequency of term i , number of documents in the collection containing i
 - N , i.e. number of documents in the collection
- TF (term frequency) and IDF (inverse document frequency) components combined multiplicatively
- Finally, in simple form

$$w_{ij} = \frac{tf_{ij}}{\sum_{k=1}^n tf_{kj}} \cdot \log\left(\frac{N}{df_i}\right)$$

Boolean retrieval model

- A document is represented as a set of keywords (index terms)
- Queries are Boolean expressions of keywords, connected by Boolean operators (AND, OR, NOT), including the use of brackets to indicate scope
 - [[Rio & Brazil] | [Hilo & Hawaii]] & hotel & !Hilton
- A document is relevant or not with respect to a query; no partial matches; no ranking
- Most systems have proximity operators (i.e. describe maximum distance between query keywords in document)
- Most systems support simple regular expressions as search terms to match spelling variants

It's always there



[Advanced Search](#)
[Preferences](#)
[Language Tools](#)



paris hilton -"paris hilton"

Find web pages that have...

all these words:

this exact wording or phrase: [tip](#)

one or more of these words: OR OR [tip](#)

But don't show pages that have...

any of these unwanted words: [tip](#)

Need more tools?

Results per page: [tip](#)

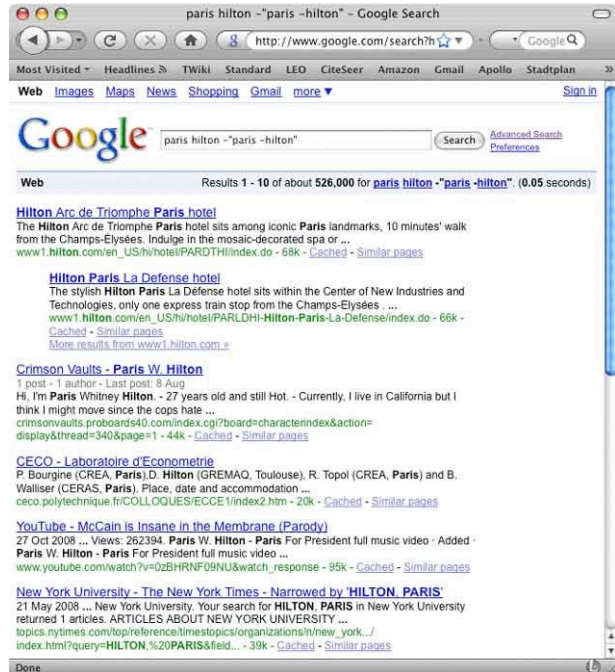
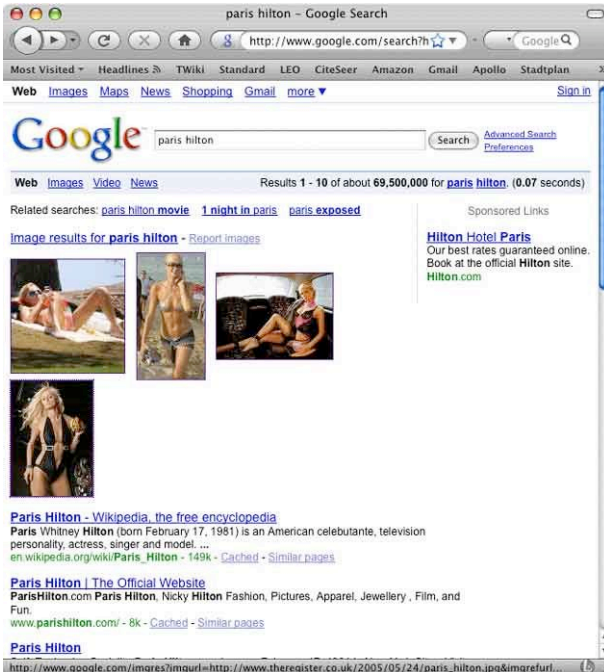
Language: [tip](#)

File type: [tip](#)

Search within a site or domain:
(e.g. youtube.com, .edu)

[+ Date, usage rights, numeric range, and more](#)

It makes a difference :-)



Vector space model

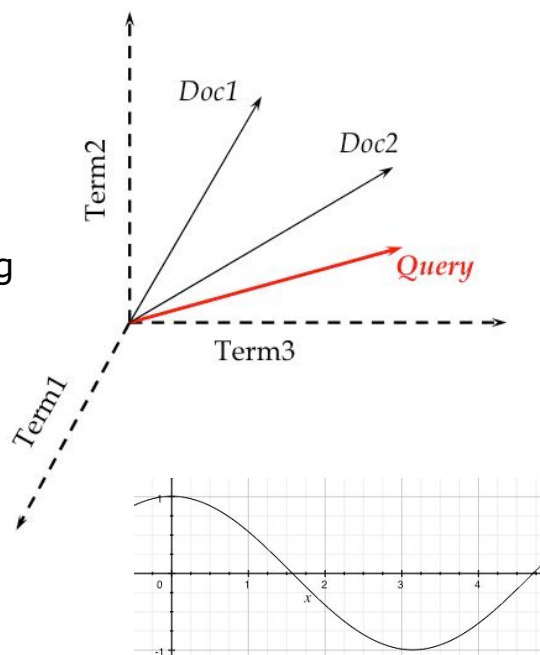
- Key idea
Everything (documents, queries, terms) is a vector in a high-dimensional space
- Formally
A vector space is defined by a set of *linearly independent* basis vectors
- Basis vectors
 - correspond to dimensions or directions in the vector space
 - determine what can be describes in the vector space
 - must be orthogonal, or linearly independent, i.e. a value along one dimension implies nothing about a value along another dimension

Vector space model

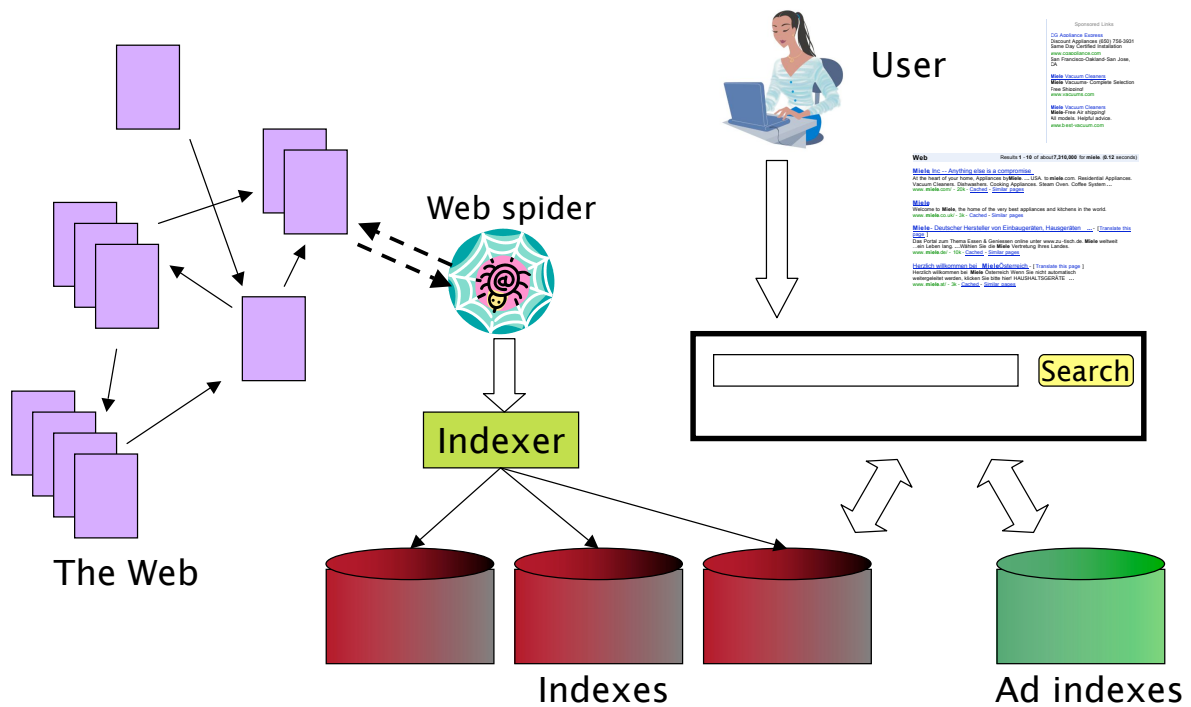
- Assume t distinct terms remain after indexing, i.e. index terms, vocabulary
- These “orthogonal” terms form a t -dimensional vector space
 $t = |\text{vocabulary}|$
- Each term i in a document (or query) j is given a real-valued weight w_{ij}
 - e.g. tf·idf, $w_{ij} = (1 + \log tf_{ij}) \log_{10}(N / df_i)$
- Both documents and queries are expressed as t -dimensional vectors
 $d_j = (w_{1j}, w_{2j}, \dots, w_{tj})$
i.e. a document (query) is represented as the sum of its term vectors

Vector space similarity

- One possibility:
Similarity is inversely related to the angle between the vectors
 $\cos(i,j) = (i \cdot j) / (|i| \cdot |j|)$
- Rank the documents by decreasing similarity to the query
- In the example, *Doc2* is the most similar to the query



Web search



User Needs

- Informational
want *to learn* about something (~40%)
 - e.g. moose
- Navigational
want *to go* to that page (~25%)
 - e.g. Kunsthistorisches Museum Wien
- Transactional
want *to do something* web-mediated (~35%)
 - like access a service - Sydney weather
 - downloads - games for the Palm Centro
 - shop - Nikon D60
- Gray areas
 - find a good hub - Car rental Lisbon
 - exploratory search - "see what's out there"

Web search users ...

- ... make ill defined queries
 - short
 - 2001: avg 2.54 terms, 80% < 3 words
 - 1998: avg 2.35 terms, 88% < 3 words
 - imprecise terms
 - sub-optimal syntax (most queries without operator)
 - low effort
- ... have wide variance in
 - needs
 - expectations
 - knowledge
 - bandwidth

Web search users ...

- ... show specific behavior
 - 85% look over one result screen only
(mostly above the fold, i.e. don't even scroll!!!)
 - 87% of queries are not modified
i.e. one query per session
 - follow links - "the scent of information"
- ... don't behave as classical IR would assume

Answering “the need behind the query”

- Semantic analysis
 - Query language determination
 - auto filtering
 - different ranking (if query in German do not return English)
 - Hard & soft (partial) matches
 - personalities (triggered on names)
 - cities (travel info, maps)
 - medical info (triggered on names and/or results)
 - stock quotes, news (triggered on stock symbol)
 - company info
- Integration of search and text analysis

Language detection - google.cz

The screenshot shows two instances of a Google search for the term "topfpflanzen".

Top instance (left): The search is performed on the Google.cz domain. The search bar contains "topfpflanzen". The results are in German, with the first result being "Imagines plantarum - Pflanzenbilder: **Topfpflanzen**". The search results are labeled "Web" and "Výsledky 1 - 10 z asi 200 000 na dotaz topfpflanzen. (0,03 sekund)".

Bottom instance (right): The search is performed on the Google.cz domain. The search bar contains "topfpflanzen". The results are in Czech, with the first result being "EUROPAGES - GENSETTER TOPFPFLANZEN AG". The search results are labeled "Web" and "Výsledky 1 - 10 z asi 61 český psaných stránek na dotaz topfpflanzen (0,03 sekund)".

"Personalities" - google.co.uk

Google the cure Search Advanced Search Preferences

Search: the web pages from the UK

Web Results 1 - 10 of about 25,700,000 for the **cure**. (0.22 seconds)

The Cure
The official **Cure** webpage. Includes news, show dates, photos, discography, video downloads, and a message board.
www.thecure.com/ - 6k - [Cached](#) - [Similar pages](#)

The Cure - Wikipedia, the free encyclopedia
This article is about the band. For the band's self-titled album, see **The Cure** (album). For other uses, see **Cure** (disambiguation). ...
en.wikipedia.org/wiki/The_Cure - 132k - [Cached](#) - [Similar pages](#)

MySpace.com - The Cure - LONDON OR LOS ANGELES, UK - Other / Other ...
MySpace music profile for **The Cure** with tour dates, songs, videos, pictures, blogs, band information, downloads and more.
www.myspace.com/thecure - 177k - [Cached](#) - [Similar pages](#)

Video results for the cure

The Cure - Just Like Heaven
3 min 12 sec
www.youtube.com

The Cure - Lovesong
3 min 27 sec
www.youtube.com

The Cure Tickets On Sale Now - Don't Miss The Cure Concerts!
The Cure Tickets available at Worldticketshop.com. Order your **The Cure** Tour Tickets online now! All **The Cure** Concert Tickets are 120% guaranteed!
www.worldticketshop.com/concerts/the_cure_tickets - 12k - [Cached](#) - [Similar pages](#)

Sponsored Links

The Cure
The Cure on eBay.
Feed your passion on eBay.co.uk!
www.ebay.co.uk

paris hilton vs hilton paris - google.com

Google paris hilton Search Advanced Search Preferences

Web News Images Video Results 1 - 10 of about 67,100,000 for **paris hilton**. (0.23 sec)

Related searches: [paris hilton movie](#) [1 night in paris](#) [paris exposed](#)

News results for paris hilton

Paris Hilton: Man Used Me for Sex and Fame - 2 Nov 2008
But sexy **Paris Hilton** snuggled up to *The Sun* and simpered that she'd been exploited by all the men in her life — hungry for fame and fortune. ...
[FOXNews - 77 related articles >](#)

Paris Hilton's 'New BFF' Candidate Performed for Girls Gone Wild ... - MarketWatch - 18 related articles >
[Flash News](#)

Paris Hilton Takes Credit for Huge Vote Turnout? - E! Online - 2 related articles >

Paris Hilton - Wikipedia, the free encyclopedia
Paris Whitney Hilton (born February 17, 1981) is an American celebratante, television personality, actress, singer and model. ...
en.wikipedia.org/wiki/Paris_Hilton - 150k - [Cached](#) - [Similar pages](#)

Paris Hilton | The Official Website
ParisHilton.com **Paris Hilton**, **Nicky Hilton** Fashion, Pictures, Apparel, Jewellery , Film, and Fun.
www.parishilton.com/ - 8k - [Cached](#) - [Similar pages](#)

Paris Hilton
Socialite **Paris Hilton** was born on February 17, 1981 in New York City... Visit IMDb for Photos, Filmography, Discussions, Blo. News, Awards, Agent, ...
www.imdb.com/name/nm0385296/ - 69k - [Cached](#) - [Similar pages](#)

Image results for paris hilton - [Report images](#)

Google hilton paris Search Advanced Search Preferences

Web Maps News Video Results 1 - 10 of about 9,420,000 for **hilton paris**. (0.31 seconds)

Hilton Hotel Paris
Hilton.com Our best rates guaranteed online. Book at the official **Hilton** site. Sponsored Link

Hilton Arc De Triomphe Paris
www1.hilton.com
51, Rue Courcelles
75008 Paris, France
+33 1 53 36 67 00
[Get directions](#)
★★★★☆ 255 reviews and more >
"Great room, so elegant. Business Lounge was terrific. Great bathroom."

Hilton Paris hotel
The **Hilton Paris** hotel sits in the shadow of the Eiffel Tower, ... From **Hilton** Guest Rooms to Suites, all **Hilton Paris** hotel rooms have wireless internet ...
www1.hilton.com/en_US/h/hotel/PARHITW-Hilton-Paris/index.do - 68k - [Cached](#) - [Similar pages](#)

[Accommodations](#) [Dining](#)
[View Maps](#) [Local Guide](#)


[More results from hilton.com >](#)

Hilton Arc de Triomphe Paris hotel
The **Hilton Arc de Triomphe Paris** hotel sits among iconic **Paris** landmarks, 10 minutes' walk from the Champs-Élysées. Indulge in the mosaic-decorated spa or ...
www1.hilton.com/en_US/h/hotel/PARDTH/index.do - 68k - [Cached](#) - [Similar pages](#)

Cities - google.com

Google [Advanced Search](#) [Preferences](#)

Web [Maps](#) [News](#) Results 1 - 10 of about 28,000,000 for **lisbon** [definition]. (0.12 seconds)



Lisbon, Portugal
maps.google.com

Start address

Remember this location

Lisbon - Wikipedia, the free encyclopedia
Lisbon (Portuguese: Lisboa, IPA: [liʒ boɐ]) is the capital and largest city of Portugal. It is also the seat of the district of **Lisbon** and capital of the ...
en.wikipedia.org/wiki/Lisbon - 210k - [Cached](#) - [Similar pages](#)

Tourist & Travel Guide | Lisbon Guide
Travel guide about **Lisbon**, Portugal with detailed information on hotels, restaurants, bars and shopping.
www.lisbon-guide.info/ - 12k - [Cached](#) - [Similar pages](#)

Lisbon Tourist Guide and Information
The capital of Portugal since its conquest from the Moors in 1147, **Lisbon** is a legendary city with over 20 centuries of History. ...
www.portugalvirtual.pt/_tourism/costadelisboa/ - 41k - [Cached](#) - [Similar pages](#)

Sponsored Links

100 Hotels in Lisbon
Book your hotel in **Lisbon** online. Good availability and great rates!
www.booking.com/Hotels-in-Lisbon

Lissabons Tourismusamt
Alle die Infos, die Sie über Europas beliebte Stadt brauchen.
www.visitlisboa.com

Hotel in Lisbon
60 Hotels - Up to 60% Discount! Charm, Design, Luxury or Family
www.FastBooking.com

Hotels in Lisbon
Book direct in 50 charming hotels. Up to 50% discount. No prepayment.
www.Hotels-Direct-Lisbon.com

Top Lisbon Hotels
The Lowest Rates in Portugal. Book the Top Portugal Hotels Now!
www.Otel.com/Portugal

Lisbon


Shopping - google.at

Google [Erweiterte Suche](#) [Einstellungen](#)

Suche: Das Web Seiten auf Deutsch Seiten aus Österreich

Web Ergebnisse 1 - 10 von ungefähr 787.000 für **buchhandlung wien**. (0,08 Sekunden)

Lokale Branchenergebnisse für buchhandlung im Umkreis von Wien



- A. **Buchhandlung am Quellenplatz - Walter Grossmann** - www.bag.co.at - 01 6041230 - [Mehr](#)
- B. **Buchhandlung Löwenherz** - www.loewenherz.at - 01 3172982 - [Mehr](#)
- C. **Leporello Buchhandlung Schöberl & Riedesser OHG** - www.leporello.at - 01 5325512 - [Mehr](#)
- D. **a.punkt Buchhandlung** - www.apunktbuch.at - 01 5328514 - [Mehr](#)
- E. **Leporello Buchhandlung Schöberl & Riedesser OHG** - www.leporello.at - 01 3198612 - [Mehr](#)
- F. **Buchhandlung und Zeitungsbüro Morawa & Co** - www.morawa-buchundmedien.at - 01 5133979 - [Mehr](#)
- G. **isbn1-meine Buchhandlung** - www.meinebuchhandlung.at - 0664 1928963 - [Mehr](#)
- H. **Wilhelm Frick Buchhandlung Gesellschaft mbH** - www.prachner.at - 01 5137364 - [Mehr](#)
- I. **Buchhandlung Bartalszky** - www.bartalszky.at - 01 3190054 - [Mehr](#)
- J. **Buchhandlung Stöhr GmbH** - www.buchhandlung-stoehr.at - 01 4061649 - [Mehr](#)

[Weitere Ergebnisse im Umkreis von Wien >](#)

Buchhandlung Herder - Wien, Zach-Buch GmbH : - **Das Herbstprogramm ...**
Buchhandlung Herder - Wien: Zach-Buch GmbH. ... Zach-Buch GmbH, Wollzeile 33, 1010 Wien. © 2007 **Buchhandlung Herder Zach-Buch GmbH**, Wollzeile 33, 1010 Wien.
www.herder.at/ - 17k - [Im Cache](#) - [Ähnliche Seiten](#)

Anzeigen

Buchladen Wien
Mit Kartenansicht, genauer Adresse, Telefonnr. & Homepage: Google Maps!
Maps.Google.at/maps

Bücher finden - AuM
Alles was das Leserherz begehrt. Spitzenauswahl, schnelle Lieferung!
www.Weltbild.at/Buecher

Jokers Buchhändler
Exklusive Bücher um 40-90% reduziert. Schnäppchen garantiert!
www.Jokers.at

Buchladen
Bücher von A-Z: Große Auswahl, kleine Preise, schnelle Lieferung.
www.bol.at

Buchhandlungen
Daten von 340.000 Unternehmen und über 3,7 Mio. Telefonteilnehmern
www.herold.at

Bücher bei Thalia.at
Geben die ganze Welt auf Ihrem...

Context transfer - google.at

Google [Erweiterte Suche](#) [Einstellungen](#)

Suche: Das Web Seiten auf Deutsch Seiten aus Österreich

Web Ergebnisse 1 - 10 von ungefähr 2.710.000 für **uggs**. (0,20 Sekunden)

Uggs - Amazon.de Anzeigen
[Amazon.de](#) Niedrige Preise, riesen Auswahl und kostenlose Lieferung ab nur 20 EUR

100% Australian Made Uggs
[www.SnugAustraliaUggBoots.com.au](#) Sale on premium quality **uggs** boots. Express delivery to Germany. Hurry!

Uggs
[www.ottoversand.at/Schuhe](#) Groß Auswahl an Markenschuhen. Jetzt bei OTTO bestellen!

Official UGG® Australia Site – UGG® Boots, Slippers, Sandals, and ... - [[Diese Seite übersetzen](#)]
 Shop the official **UGG®** Australia site for **UGG®** boots for women, men and kids. Also shop our **UGG®** slippers collections and new **UGG®** women's espadrilles, ...
[www.uggaustralia.com/](#) - 21k - [Im Cache](#) - [Ähnliche Seiten](#)

[Women's UGG® Boots](#) [Official UGG® Australia Brand Site](#)
[Shop Kids Boots, Casuals ...](#) [Big Kids](#)
[Canada](#) [Casuals](#)

[Weitere Ergebnisse von uggaustralia.com »](#)

UGG Boots, Uggs - Free Shipping - [[Diese Seite übersetzen](#)]
Ugg boots, **uggs**. Only the genuine **UGG** Australia brand are actually **UGGS**. Buy direct from the original online retailer. Free shipping. no sales tax.

Uggs - Huge Sale on Boots Anzeigen
 Up to 70% Off, Hurry Sale Ends Soon
 Australian Real Sheepskin Boots
[www.UggStopSale.com.au](#)

Uggs
 Ob Sandalen, Sneakers, Stiefel oder Ballerinas - Schuhe bei Universal!
[www.universal.at/Schuhe](#)

Ugg Boots bei BuyVIP
 Finden Sie Ihre Lieblingsmarken im Shopping-Club zu Outlet-Preisen!
[www.buyvip.com](#)

Solid Edge ST (V21)
 100x schnelleres Arbeiten im CAD mit "Synchronous Technology"
[www.pbu-cad.at](#)

Uggs
 Riesenauswahl zu Niedrigpreisen. Hunderte von Angeboten: bei eBay!
[www.ebay.at](#)

Context transfer - google.at

Uggs - Amazon.de
[Amazon.de](#) Niedrige Preise, riesen Auswahl und kost 20 EUR

100% Australian Made Uggs
[www.SnugAustraliaUggBoots.com.au](#) Sale on premium quality **uggs** boots. Express delivery to Germany. Hurry!

Uggs
[www.ottoversand.at/Schuhe](#) Groß Auswahl an Markenschuhen. Jetzt bei OTTO bestellen!

amazon.de Hallo, Dieter Merkl. Wir haben [Empfehlungen](#) für Sie. ([Ausloggen](#))

Dieters Amazon.de [Sonderangebote](#) [Wunschzettel](#) [Gutscheine](#) [Geschenke](#)

Alle Kategorien ansehen


Amazon.de [Gutscheine](#) [Hollywood](#) [Bestseller](#) [Sonderangebote](#) [Jetzt Verkauft](#) [Disney](#)


Abteilung
Beliebige Abteilung
 Englische Bücher (1.592)
 Bücher (379)
 Schuhe & Handtaschen (56)
 Sport & Freizeit (12)
 Spielzeug (11)
 Musik (3)
 Garten & Freizeit (1)
 Zeitschriften (1)


Versandart
Jede Versandoption
 Artikel, die bei Amazon.de auf Lager sind (548)

Lieblingslisten

"uggs"
 Ergebnisse aus allen Kategorien [« Zurück](#) | **1** | [Weiter »](#)

1.  **UGG Women's Classic Short Chocolate, braun** von L
[Neu kaufen: EUR 199,00](#)
Kostenlose Lieferung möglich.
Schuhe & Handtaschen: Alle 56 Artikel ansehen

2.  **UGG UGG Beacon 5525-07W, Herren Stiefel, beige,**
[Neu kaufen: EUR 249,00 EUR 159,90](#)
Kostenlose Lieferung möglich.
Schuhe & Handtaschen: Alle 56 Artikel ansehen

3.  **UGG UGG Broome - Leather 5511-07W, Damen Stiefel**
[Neu kaufen: EUR 429,00 EUR 199,90](#)
Kostenlose Lieferung möglich.

No transfer - google.at

The image shows a screenshot of the OTTO website. On the left, there is a yellow box containing search results for 'Uggs'. A red arrow points from the link 'www.ottoversand.at/Schuhe' in this box to the 'Schuhe' category in the main navigation menu. The main website header includes the OTTO logo, a search bar, and various navigation links. The main content area features a large pink banner for 'Schuhe' with three sub-sections: 'DAMEN', 'HERREN', and 'KINDER & TEENS'. The footer contains the logo for 'eC electronic commerce group' and 'TU VIENNA'.

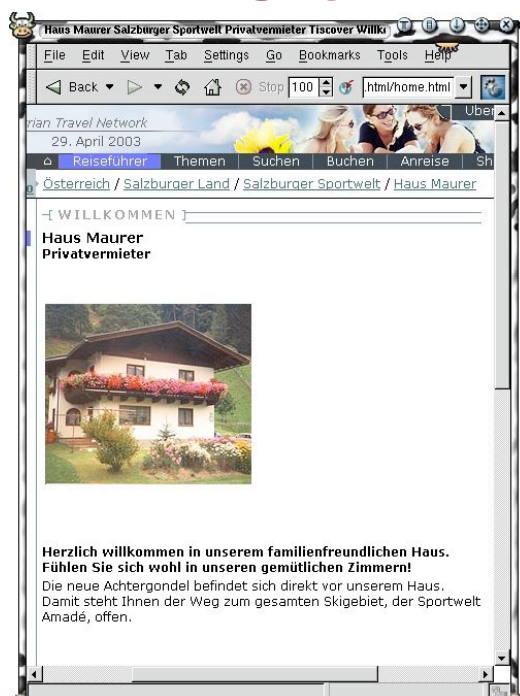
Where to go from here?

- Text mining
- Concept discovery

Text mining - Ontology

- Ontology enhancement
 - clustering of domain-related terms occurring in free-form text descriptions according to their similarity (two-dimensional map display)
 - extraction of words/concepts from free-form text descriptions that are important for specific geographic regions

Text mining - Ontology



Text mining - Ontology

- very different styles, texts are written by the accommodation providers themselves
- accommodation descriptions are dominated by enumerations of services and facilities
- semantically similar words are located close to each other regarding their position in the text
- similar structure can be found in other product descriptions

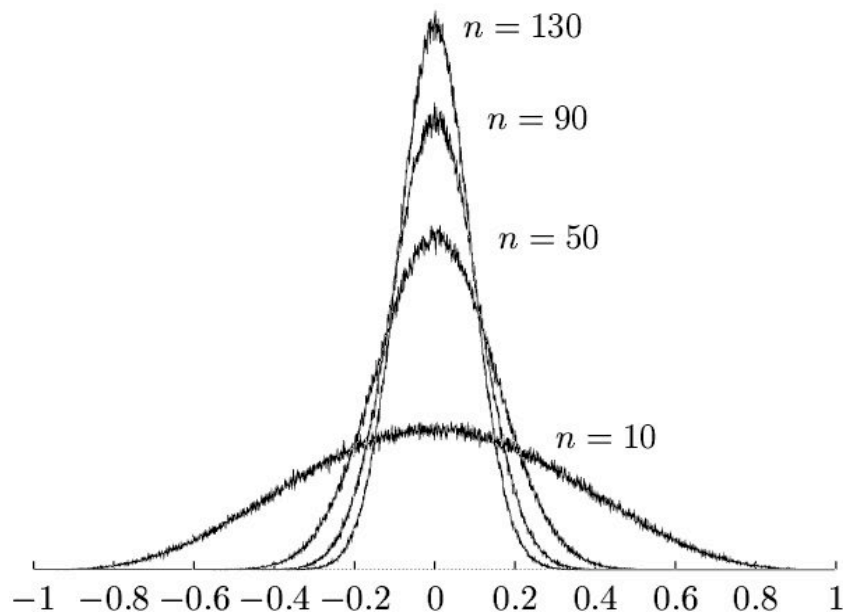
Text mining - Ontology

- Preprocessing
- remove words other than nouns and proper names to avoid primary clustering according to word classes
 - select words starting with capital letter in german texts
 - part-of-speech taggers possible for other languages

Text mining - Ontology

- Random Mapping
- “true” independence of vector representation is computationally not feasible
- assign n -dimensional random vector to each word ($n=90$)
- random values of vector components are drawn from a uniform distribution => quasi-orthogonal vectors
- sufficient independence of vectors to avoid unwanted distortions

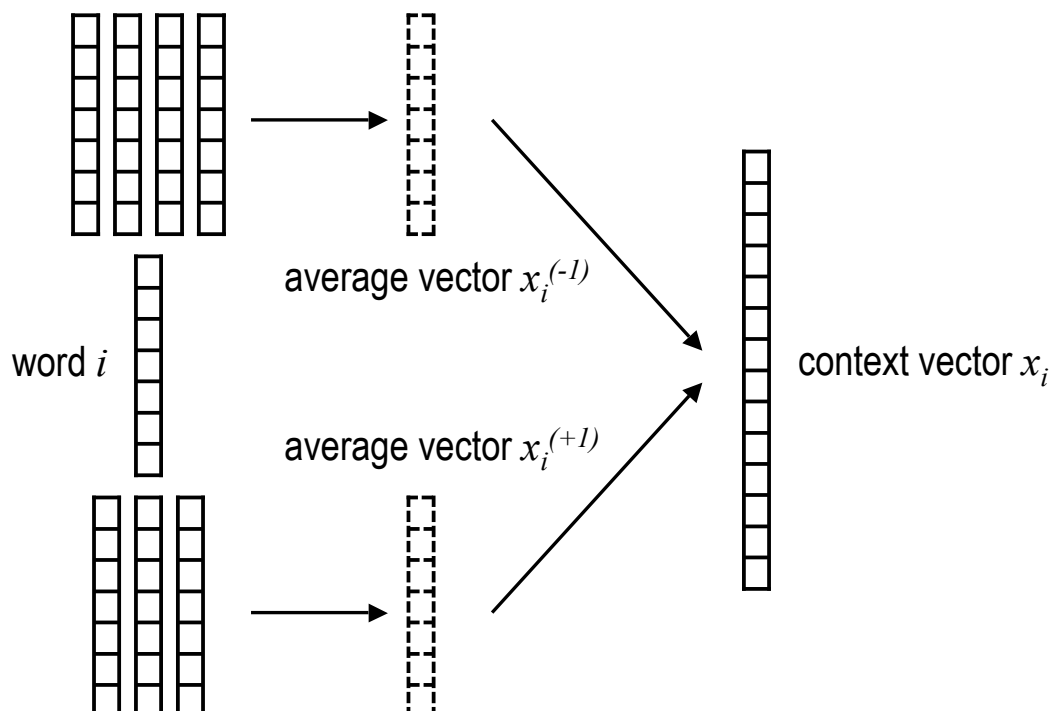
Text mining - Ontology



Text mining - Ontology

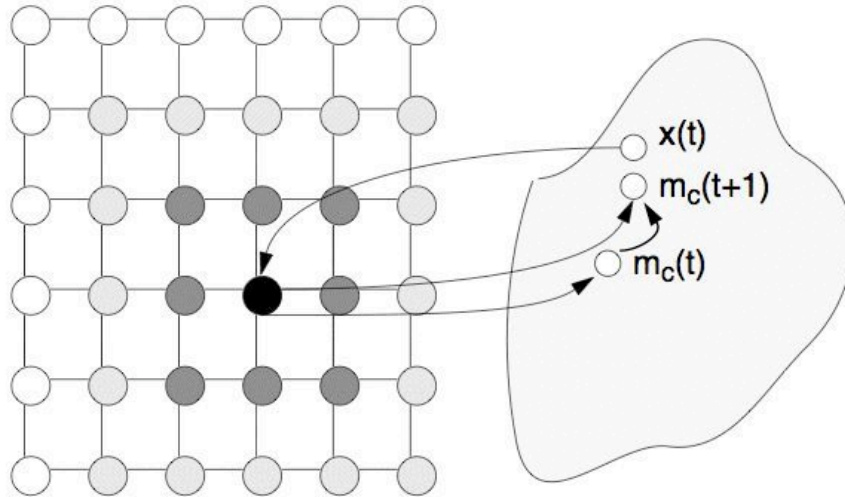
- a list of terms at different displacements is created for each word (e.g. all directly preceding terms at position -1)
- average vectors are calculated => average context
- average context vectors are concatenated to create a vector description of a word determined by its surrounding words
- example: Skifahren
 - words at displacement -1: Langlaufen, Rodeln, Pulverschnee, Winter, ...

Text mining - Ontology

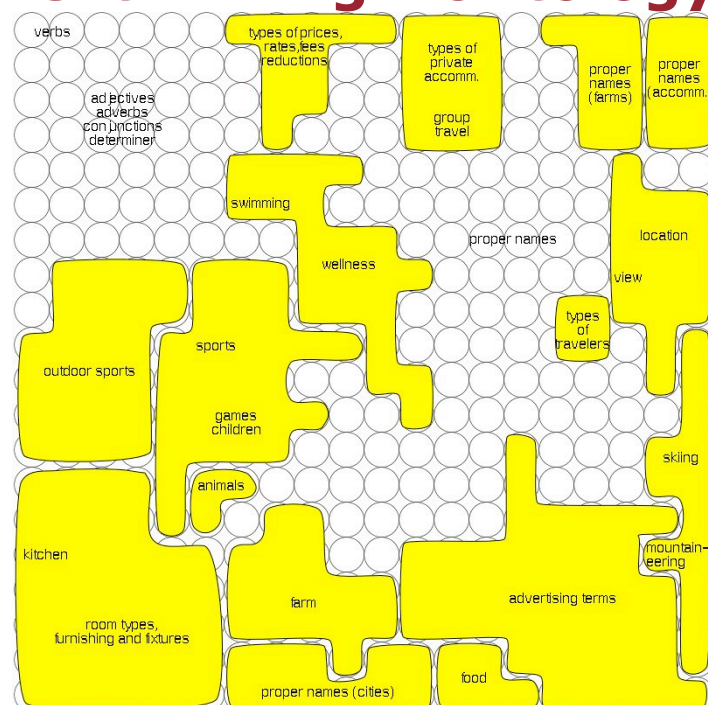


Text mining - Ontology

- Self-organizing map



Text mining - Ontology



Text mining - Ontology

- Detail - lower left corner

toilette	kuechenblock	kochnische	bad
suedbalkon	essecke	wanne	stockbett
wohnbereich	wohnkueche	sofa	doppelzimmern
diele	couch	badewanne	doppelbettzimmer
elektroheizung	schlafgelegenheit	waschraum	usche
garderobe	ausziehcouch	doppelbett	schlafraeume
doppelwaschbecken	vorraum	schlafmoeglichkeiten	zimmerausstattung
wohnkuechen	stockbetten	hotelzimmer	dreibettzimmer
wc	kuechenzeile	essraum	wohnschlafrum
bidet	wohnzimmer	kochecke	schlafzimmer
	essplatz	duschen	zimmer
	esszimmer	kinderzimmer	fliesswasser
	doppelcouch	schlafrum	einbettzimmer
	wohnraum	wohnschlafzimmer	komfortzimmer
	flur	badezimmer	doppelschlafzimmer
		wohnstube	schlafraeumen
			gaestezimmer

Text mining - Ontology

- stunning diversity of terms describing very similar concepts
- example: terms describing recreational facilities having in common that the vacationer sojourns in a closed room with well-tempered atmosphere:
 - Sauna, Tepidarium, Biosauna, Kräutersauna, Finnische Sauna, Dampfsauna, Dampfbad, Thermarium, Infrarotkabine, ...

Text mining - Geography

- rank terms according to their importance for a specific geographic region
- based on occurrence frequencies in text documents
- different granularities
 - federal state
 - region
 - city
 - ...

Text mining - Geography

- rf_{ik} ... number of documents related to a region k where term i occurs
- N_k ... number of documents related to a region k

$$w_{ik} = \frac{rf_{ik}}{N_k} \times \frac{1}{\sum_l \frac{rf_{il}}{N_l}}$$

Text mining - Geography

- $w_{ik}=1$, if term i occurs only in documents of region k and nowhere else
- if $w_{ik}<1$:
 - w_{ik} as well as the standard deviation of a term's weights indicates its distribution and can be used as a measure for ranking
 - stop words (and, the, ...) and general terms (urlaub, gast, ...) are evenly distributed => low standard deviation

Text mining - Geography

- Example - Vienna

rank	term	rank	term	rank	term
1	stephansdom	11	mariahilferstraße	21	biedermeierstil
2	ringstraße	12	einkaufsstraßen	22	westbahnhofes
3	staatsoper	13	burgtheater	23	walzer
4	stephansplatz	14	air	24	vollklimatisierten
5	mariahilfer	15	u-bahnstation	25	uno
6	westbahnhof	16	riesenrad	26	spittelberg
7	schönbrunn	17	raimundtheater	27	parlament
8	ringstrasse	18	kärntnerstraße	28	opernkarten
9	prater	19	donauinsel	29	altwiener
10	wien-aufenthalt	20	museumsquartier	30	wienerberg

Text mining - Geography

- Example - Crossing borders

Terms	Federal States								
	Vie	Low. A	Upp. A	St	Bgl	Sbg	Car	Tyr	Vbg
Salzkammergut	0	0	0.8	0.14	0	0.06	0	0	0
Salzkammergutes	0	0	0.76	0.11	0	0.13	0	0	0
Salzkammergutseen	0	0	0.89	0	0	0.11	0	0	0
Arlberg	0	0	0	0	0	0	0	0.11	0.89
Arlberger	0	0	0	0	0	0	0	0.15	0.85
Arlbergs	0	0	0	0	0	0	0	0.02	0.98
Thermenland	0	0	0	0.88	0.12	0	0	0	0
Thermenregion	0	0.13	0.16	0.35	0.36	0	0	0	0
Thermenhotel	0	0	0.2	0.62	0.18	0	0	0	0

Text mining - Clustering

- Goal: Grouping of "similar" documents, i.e. documents covering a "similar" topic
- "Bag of Words" approach for indexing
- tf*idf term weights
- Self-organizing map for clustering
- Results in a "map" of the document space
-> "similar" documents are shown in spatial proximity on the map
- Examples
 - TIME articles from the 1960s
 - Country descriptions from the CIA World Factbook

Netscape - [Time Magazine, 10 x 15 SOM]

File Edit View Go Bookmarks Options Directory
Window Help

The Migration Collection

indian, negotiation, settlement, delhi, nehru, india, round, pakistan

soviet, moscow, nuclear, khrushchev, negotiation, treaty, berlin, west, russia, agreement, pact, undergo, test

austrian, people, conservative, socialist, coalition, argument, austria, ministry

souvanna, geneva, red, laotian, pathet, viet, vientian, plain, laos, jarr, kong, neutrality

viet, saigon, catholic, religious, priest, blame, monk, diem, buddhist, quang

viet, saigon, vietnamese, crusage, dinh, diem, monk, buddha, barricade, blame, thuc, cong, catholic, religious, buddhist

malaysia, malayan, brunei, federation, borneo, singapore, abdul, malaya, indonesia, philippines, rahman, tunku

nuclear, multilateral, crew, european, manned, all, polaris, fleet, britain, europe, submarine, deterrent, contribution, ship, nato

deterrent, nato, missile, all, nuclear, submarine, multilateral, nassau, skybolt, gaulle, france, polaris

common, ivanov, debate, tory, christine, profumo, keeler, ward, macmillan, wilson

girl, ivanov, keeler, profumo, christine, ward

Document: Done

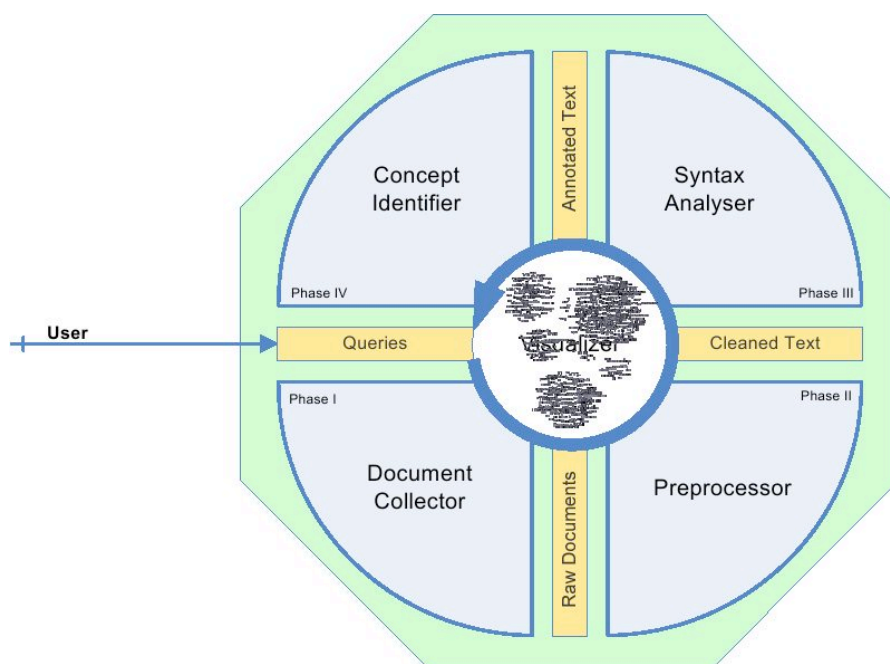
south, viet, vietnamese, dinh, lodge, cong, saigon, diem, minh, monk, pagoda, buddhist

Christmas Island Cocos Islands	Norfolk Island Saint Pierre North	Guam Mariana Islands	American Samoa	Marshall Islands Palau	Papua New Guinea Solomon Islands	Tonga Western Samoa	Sao Tome	Chad Mali Niger	Burundi Rwanda Uganda
Cook Islands Niue Tokelau Tuvalu	Wallis	Guadeloupe Martinique	Aruba Puerto Rico Virgin Islands	Antigua Grenada Saint Kitts Saint Lucia Saint Vincent	Kiribati Mauritius Nauru Seychelles Vanuatu	Comoros Maldives	Cape Verde Djibouti Equatorial Guinea	Burkina Faso Central African Rep. Guinea Guinea Bissau	Gambia Sierra Leone
Anguilla Falkland Islands Saint Helena	Mayotte New Caledonia	French Guiana French Polynesia	Hong Kong Netherlands	Barbados	Belize	Bhutan Nepal	Angola Madagascar Mozambique Nigeria	Botswana Lesotho Malawi Swaziland Zambia Zimbabwe	Senegal
British Virgin Islands Montserrat Pitcairn Islands Turks Islands	Guernsey Jersey	Macau Reunion	Malta	Bahamas Jamaica	Guyana Suriname Trinidad	Afghanistan Cambodia Laos	Namibia	Kenya Tanzania	Cameroon Gabon Ghana
Bermuda Cayman Islands Gibraltar Isle of Man	Faroe Islands Greenland	Andorra San Marino Vatican	South Africa	India Pakistan	Bangladesh Burma Thailand	Brunei Cyprus Fiji Liberia	Ethiopia Somalia	Mauritania Zaire	Benin Congo Ivory Coast Togo
South Georgia Svalbard World	Gaza Strip West Bank	Liechtenstein Luxembourg Monaco	Italy	Greece Ireland Turkey	Singapore South Korea Sri Lanka	Iran Vietnam	Bahrain Kuwait Oman Qatar South Yemen United Arab Emirates	Lebanon North Yemen Saudi Arabia	Libya
Jan Mayen	Iraq-SA Neut. Zone	Iceland Norway	Austria Belgium France German Fed. Rep. Spain Switzerland	Canada Portugal United Kingdom	Australia New Zealand	China Taiwan	Sudan	Israel Jordan Syria	Algeria Morocco Tunisia
Ashmore Islands Coral Sea Islands Heard Island	Paracel Islands Spratly Islands	Denmark Finland Sweden	Japan	United States	Soviet Union Yugoslavia	Egypt Iraq	Indonesia Malaysia Philippines	Honduras	
Bassas da India Clipperton Island Europa Island Glorioso Islands Juan de Nova Island Tromelin Island	Navassa Island Wake Island	Antarctica	Albania	Mongolia North Korea	Panama	Paraguay Venezuela	Costa Rica El Salvador Guatemala Haiti Mexico Nicaragua		
Bouvet Island French Antarctic Lands Jarvis Island Kingman Reef Palmyra Atoll	Baker Island Howland Isl. Midway Islands	Johnston Atoll Midway Islands	Arctic Ocean Atlantic Ocean	Indian Ocean Pacific Ocean	Bulgaria Czechoslovakia German Dem. Rep. Hungary Poland Romania	Cuba	Bolivia Peru	Argentina Brazil Chile Colombia Dominican Rep. Ecuador Uruguay	

Concept discovery

- Motivation: create research instrument that
 - transcends traditional, keyword-based search engines by uncovering different (context-sensitive) meanings of concepts and their relations to other concepts
 - uses the Web as information source being independent of manually created annotations
- 4-phase process, 2 iterations
- current prototype uses Google, Altavista, Yahoo!
- *is-a* relations

Concept discovery



Concept discovery - Iteration 1

Document Collector (Phase I)

- initial query term(s) provided by user
e.g. Multiple Sclerosis
- creation of search engine-dependent queries
e.g. google: "Multiple Sclerosis is (a OR an OR the)"
- send queries to search engines
- collect lists of URLs and merge
- retrieve documents

that's important
to find relations describing
what something is rather
than *how*

Concept discovery - Iteration 1

Preprocessor (Phase II)

- cleaning of documents, conversion to plain text (currently
PDF, RTF, HTML)
- HTML: improve punctuation based on tags

Concept discovery - Iteration 1

Syntax Analyzer (Phase III)

- sentence splitter
- selection of relevant (matching) sentences
- Part-of-Speech tagging and noun phrase chunking

Concept discovery - Iteration 1

Concept Identifier (Phase IV)

- select first noun phrase **after** verb
- add concept to graph, if not already present

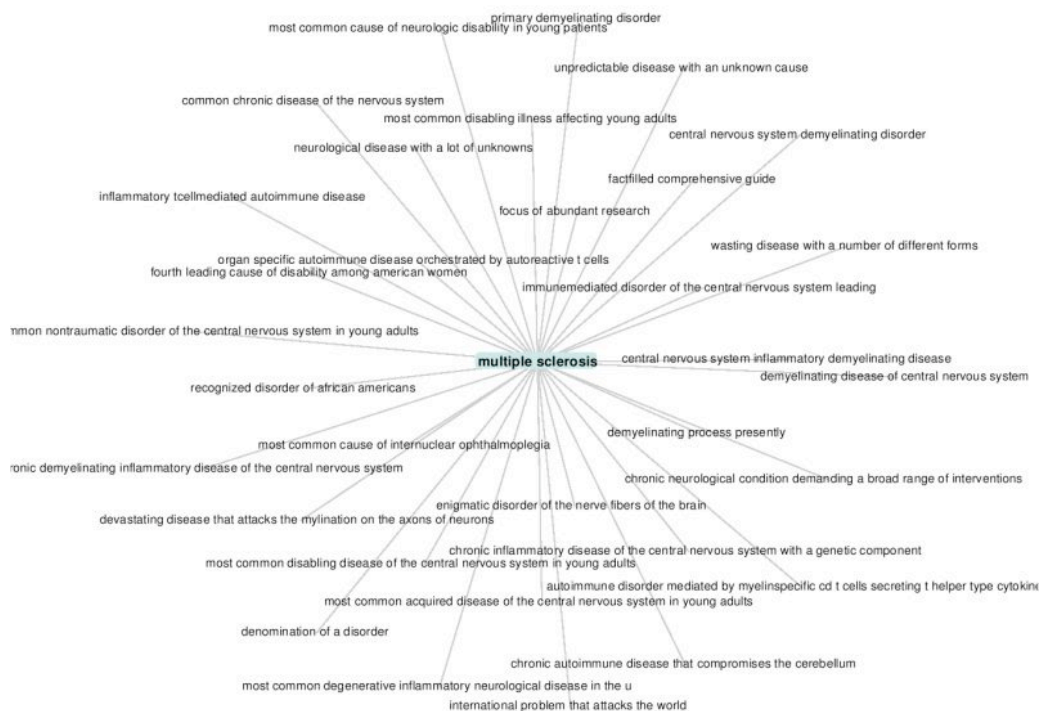
Concept discovery - Iteration 2

- for each identified concept from the first iteration, apply Phase I-IV with two important differences:
 - query generation in phase I:
"is (a OR an OR the) <concept name>"
 - concept selection in Phase IV: select first noun phrase **before** the verb

Concept discovery

- Example: Microsoft Windows
- Iteration 1
 - "Microsoft Windows is (a OR an OR the)"
 - Result of Iteration 1: e.g. operating system
- Iteration 2
 - "is (a OR an OR the) operating system"
 - Result of Iteration 2: e.g. Linux, MacOS, Plan 9, CentOS,
...

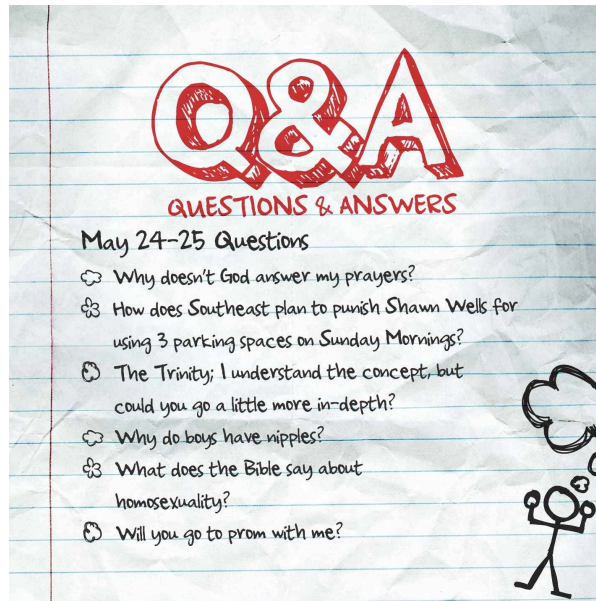
Example: Multiple Sclerosis



Interesse noch nicht komplett vergangen?

- Zwei recht dicke und ganz feine Bücher (mehr oder weniger) zum Thema (natürlich viel ausführlicher)
- C. D. Manning & H. Schütze: *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA. 2000.
- C. D. Manning, P. Raghavan, H. Schütze: *Introduction to Information Retrieval*. Cambridge University Press. New York, NY. 2008.
Available online at <http://www.informationretrieval.org/>

Gibt's Fragen?



**Remember, we live in a world of
digital divide :-)**

