Information Retrieval

PS Einführung in die Computerlinguistik
SE aus Artificial Intelligence

Dieter Merkl
dieter.merkl@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

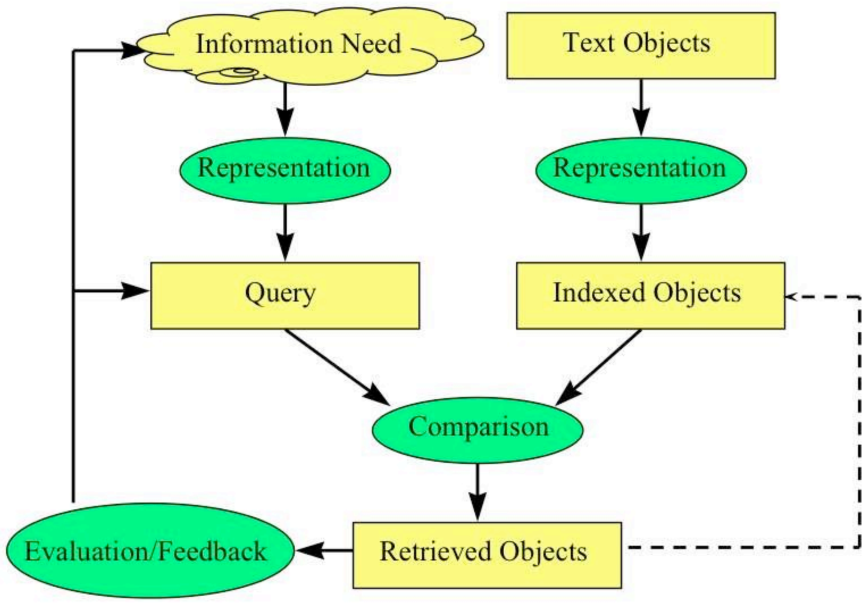
Guess what's this about?

- 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



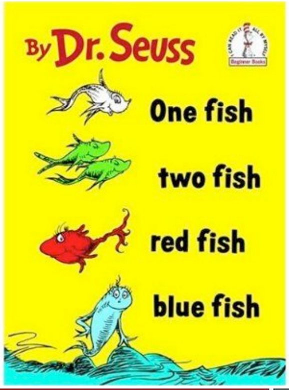
CMPSCI 646

Example: Small document

- D = {one fish, two fish, red fish, blue fish, black fish, blue fish, old fish, new fish}

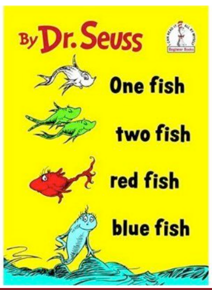
- len(D) = 16

- P(fish|D) = 8/16 = 0.5
- P(blue|D) = 2/16 = 0.125
- P(one|D) = 1/16 = 0.0625
- …
- P(eggs|D) = 0/16 = 0



Example: Three small documents

- D1 = {This one, I think, is called a Yink. He likes to wink, he likes to drink.}
- D2 = {He likes to drink, and drink, and drink. The thing he likes to drink is ink.}
- D3 = {The ink he likes to drink is pink. He likes to wink and drink pink ink.}

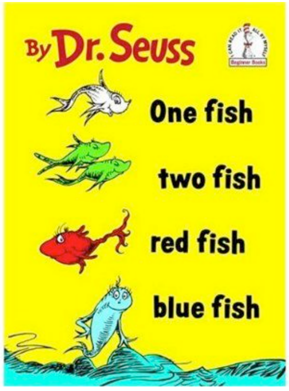


- Query "drink"
  - P(drink|D1} = 1/16
  - P(drink|D2) = 4/16
  - P(drink|D3) = 2/16
- Query "pink ink"
  - P(pink ink|D1) = 0 · 0 = 0
  - P(pink ink|D2) = 0 · 1/16 = 0
  - P(pink ink|D3) = 2/16 · 2/16 ≈ 0.016
- Query "wink drink"
  - P(wink drink|D1) = 1/16 · 1/16 ≈ 0.004
  - P(wink drink|D2) = 0
  - P(wink drink|D3) = 1/16 · 2/16 ≈ 0.008

ec electronic commerce group
Institute of Software Technology and Interactive Systems
TU VIENNA

Danke für den Hinweis während des Vortrags!

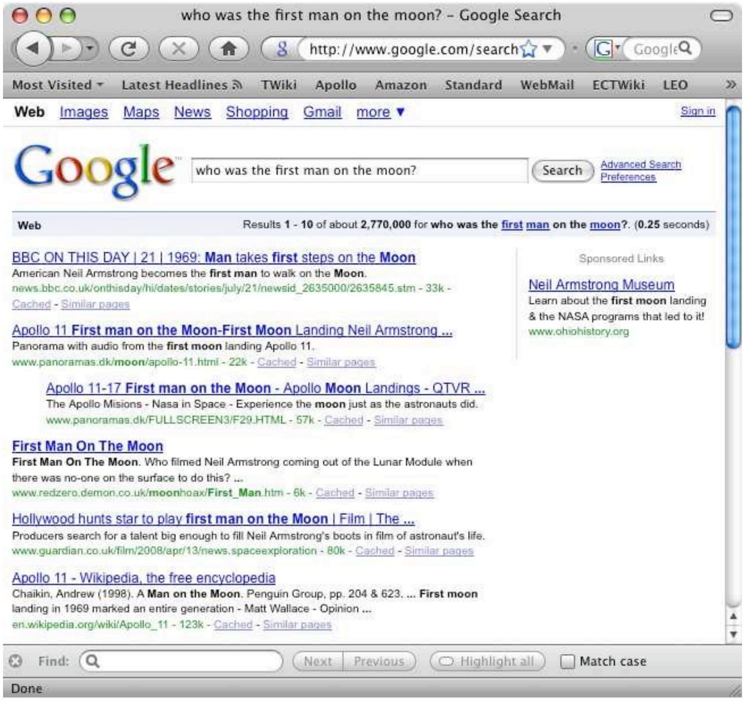
- Die Stelle ist wohl wirklich aus
  "One fish, two fish, red fish, blue fish" …
  … und nicht wie fälschlicherweise behauptet aus
  "Green eggs and ham"
  - This one, I think, is called a Yink.
  - He likes to wink,
  - he likes to drink.
  - He likes to drink, and drink, and drink.
  - The thing he 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"

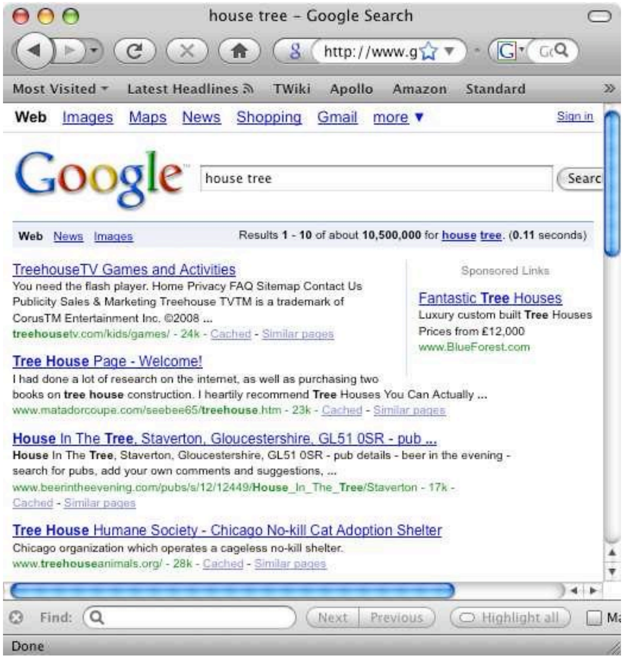
Who was the first man on the moon?

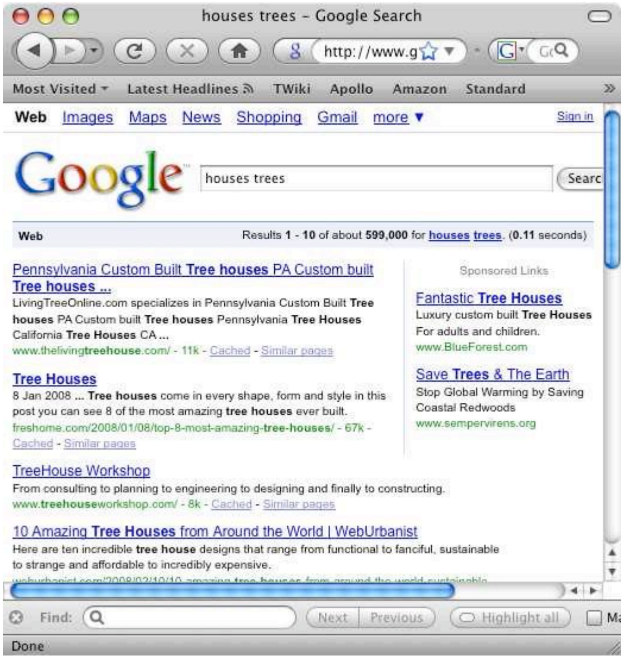


Basic automatic indexing

- 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





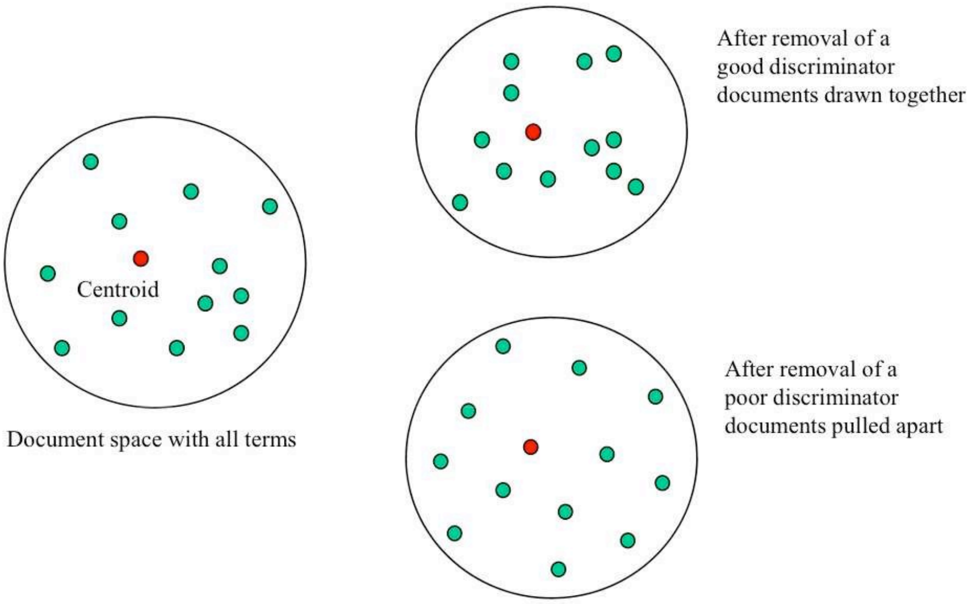
Indexing models

- 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	marrow	Buddhist
flutter	Amyloidosis	Diem
jet	Lymphostasis	Lao
cone	Hepatitis	Arab
separate	Hela	Viet
shell	antigan	Kurd
yaw	chromosome	Wilson
nozzle	irradiate	Baath
transit	tumor	Park
degree	virus	Nenni
	Worst Discriminators	
equate	clinic	work
theo	children	lead
bound	act	Red
effect	high	minister
solution	develop	nation
method	treat	party
press	increase	commune
result	result	U.S.
number	cell	govern
flow	patient	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

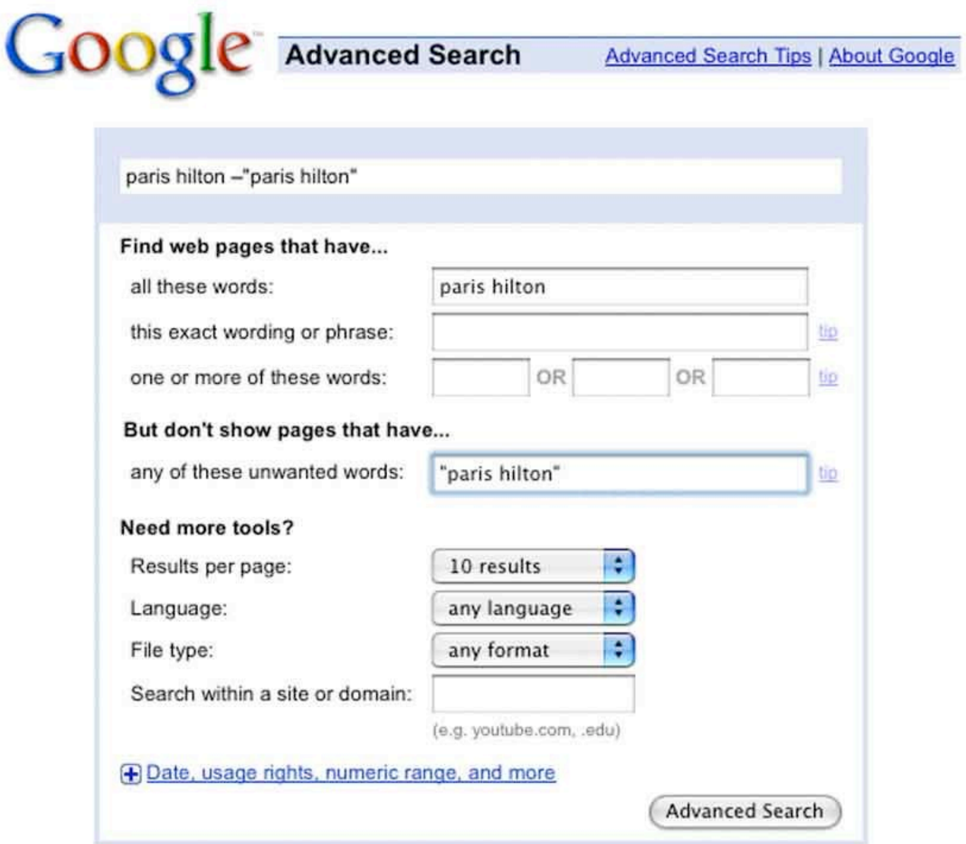
- We're there, at last :-)
- Notation
  - df_i, i.e. document frequency of term i, number of documents in the collection containing i
  - N, i.e. number of documents in the collection
- TF (term frequency) and IDF (inverse document frequency) components combined multiplicatively
- Finally, in simple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

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

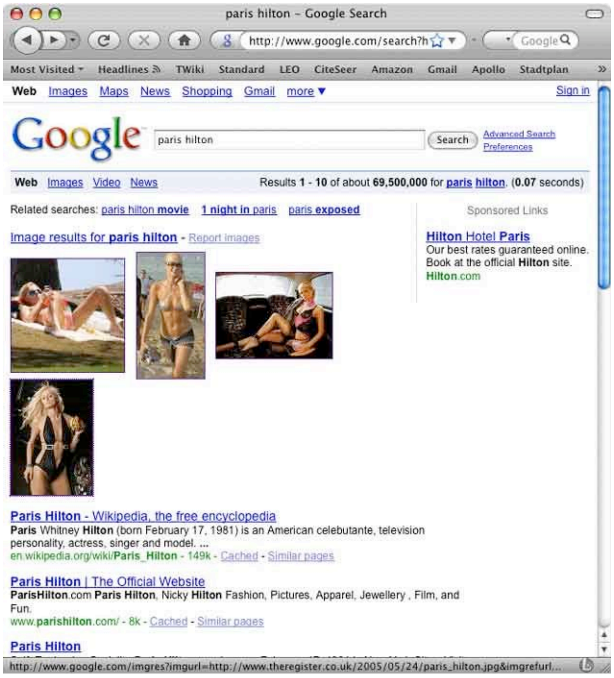
Boolean retrieval model

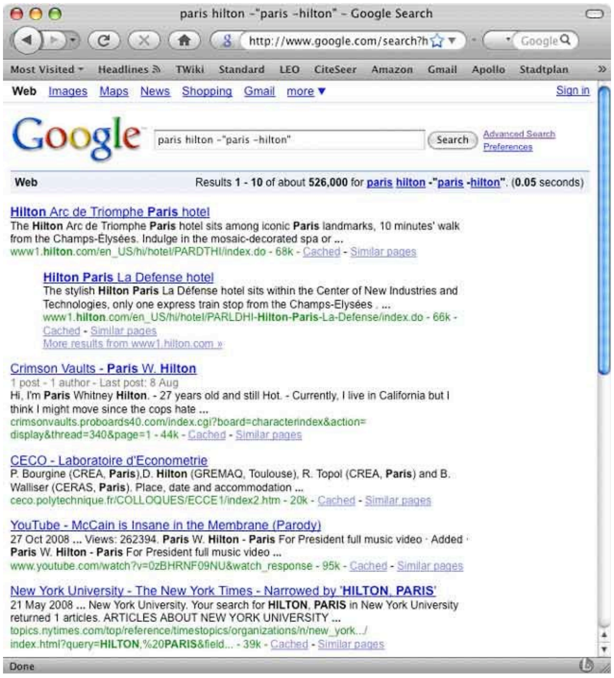
- A document is represented as a set of keywords (index terms)
- Queries are Boolean expressions of keywords, connected by Boolean operators (AND, OR, NOT), including the use of brackets to indicate scope
  - [[Rio & Brazil] | [Hilo & Hawaii]] & hotel & !Hilton
- A document is relevant or not with respect 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



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

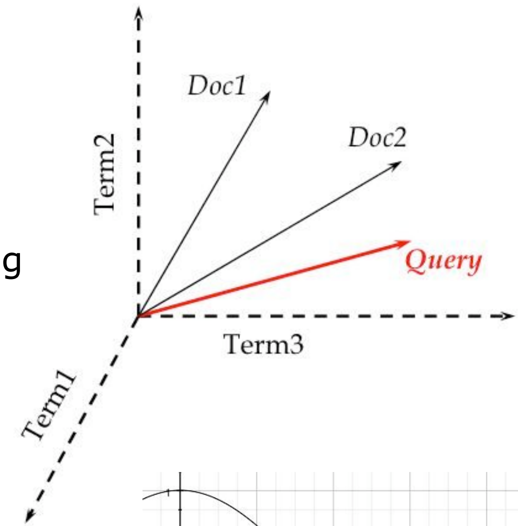
ec electronic commerce group
Institute of Software Technology and Interactive Systems
TU VIENNA

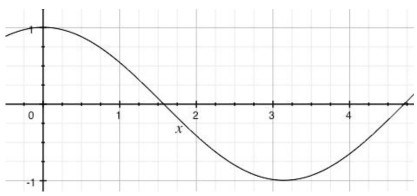
Vector space model

- Assume t distinct terms remain after indexing, i.e. index terms, vocabulary
- These "orthogonal" terms form a t-dimensional vector space
  t = | 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
  $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$
  i.e. a document (query) is represented as the sum of its term vectors

Vector space similarity

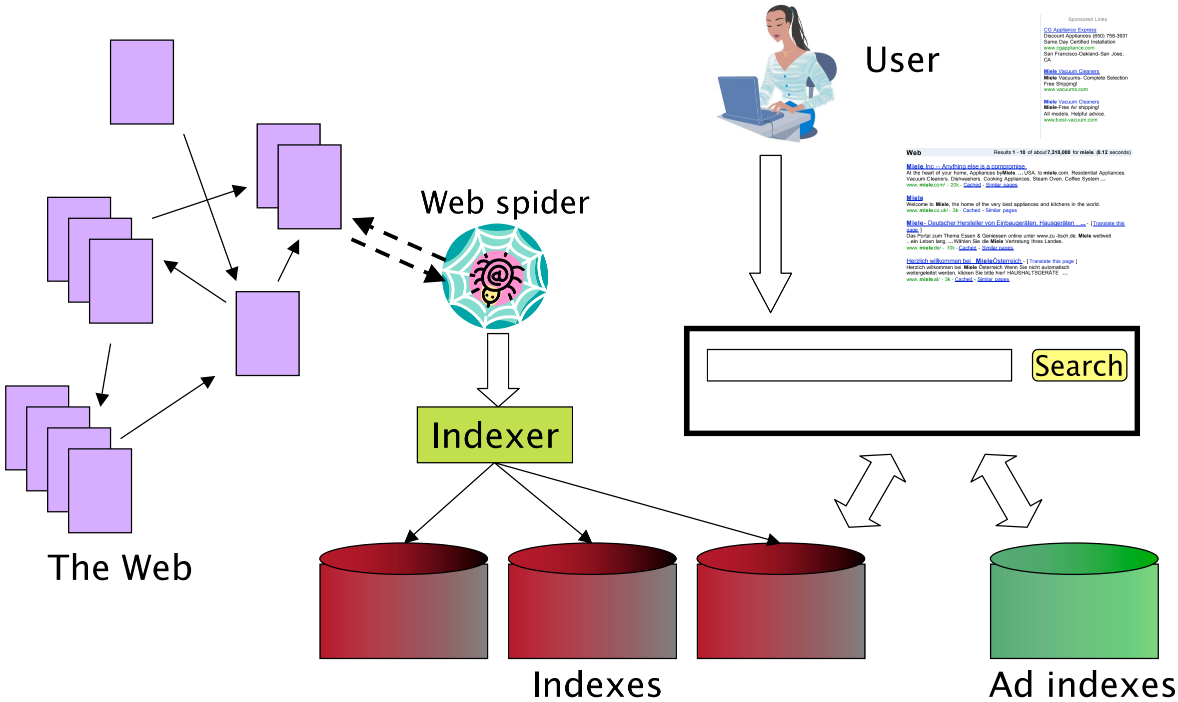
- One possibility:
  Similarity is inversely related to the angle between the vectors
  $cos(i,j) = (i*j)/(|i|*|j|)$
- Rank the documents by decreasing similarity to the query
- In the example, *Doc2* is the most similar to the query





ec electronic commerce group
Institute of Software Technology and Interactive Systems
TU VIENNA

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 ...

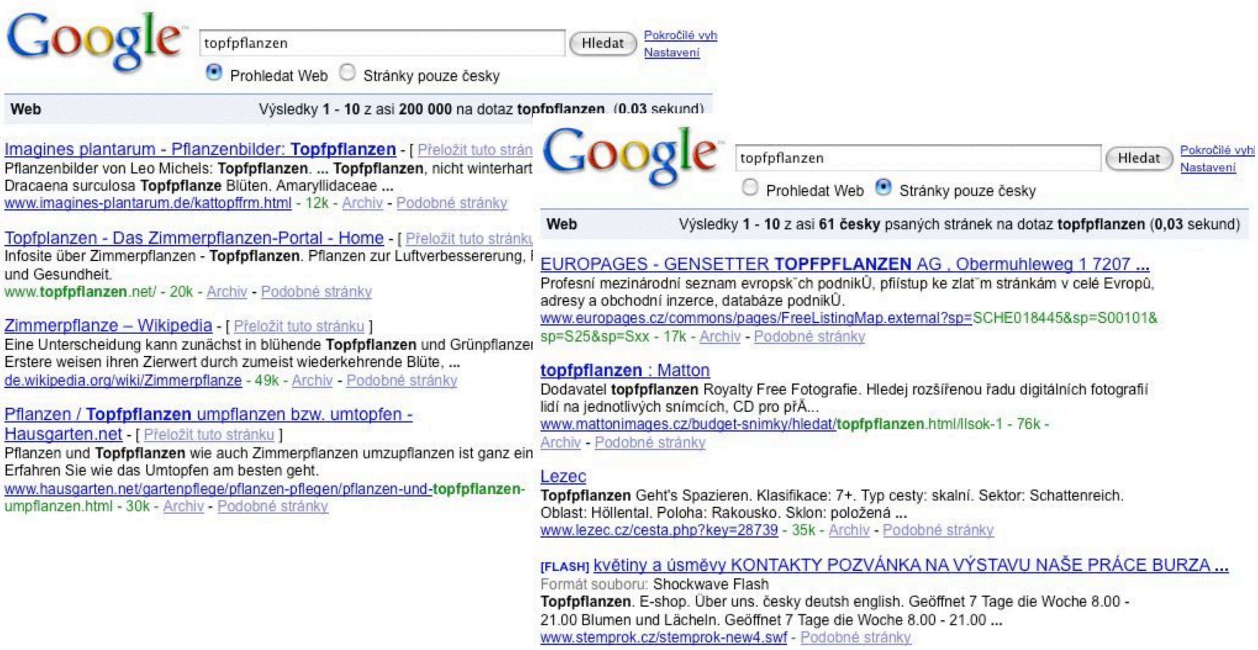
- ... show specific behavior
  - 85% look over one result screen only
    (mostly above the fold, i.e. don't even scroll!!!)
  - 87% of queries are not modified
    i.e. one query per session
  - follow links - "the scent of information"
- ... don't behave as classical IR would assume

ec electronic commerce group
Institute of Software Technology and Interactive Systems
TU VIENNA

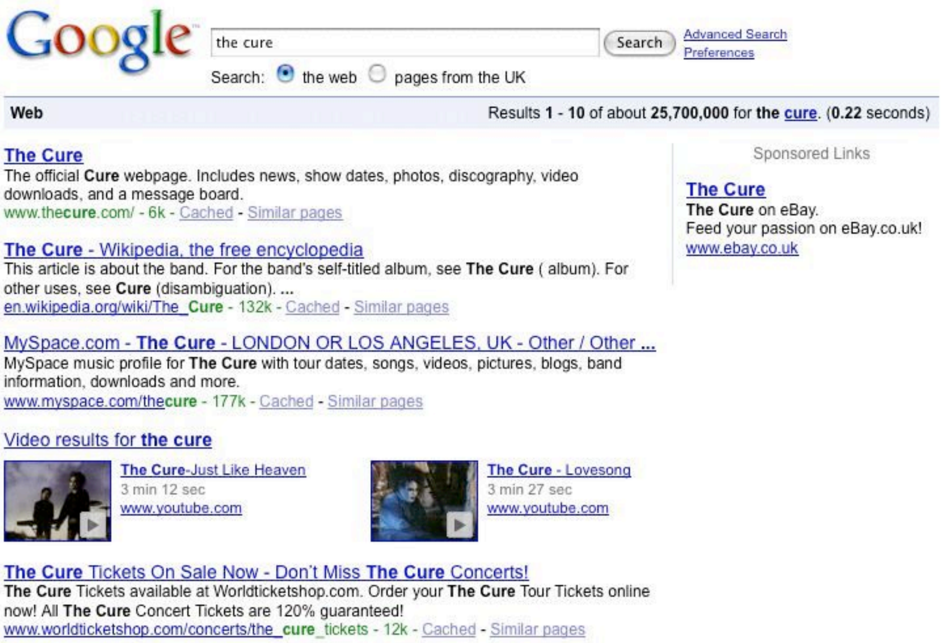
Answering
“the need behind the query”

- 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

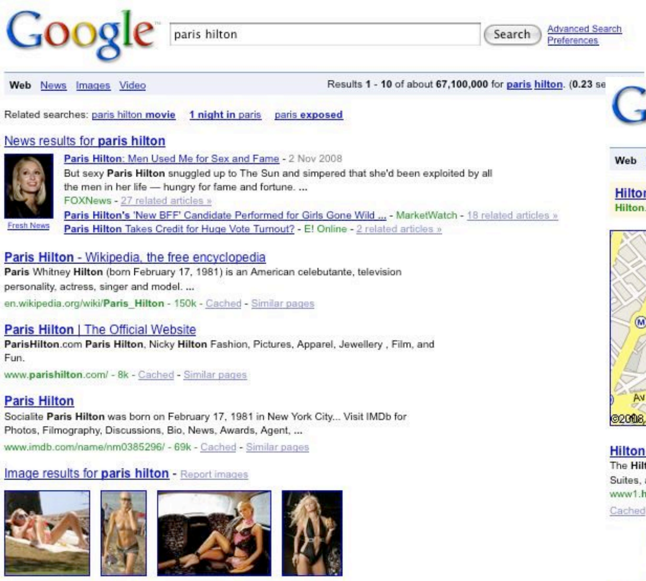
Language detection - google.cz

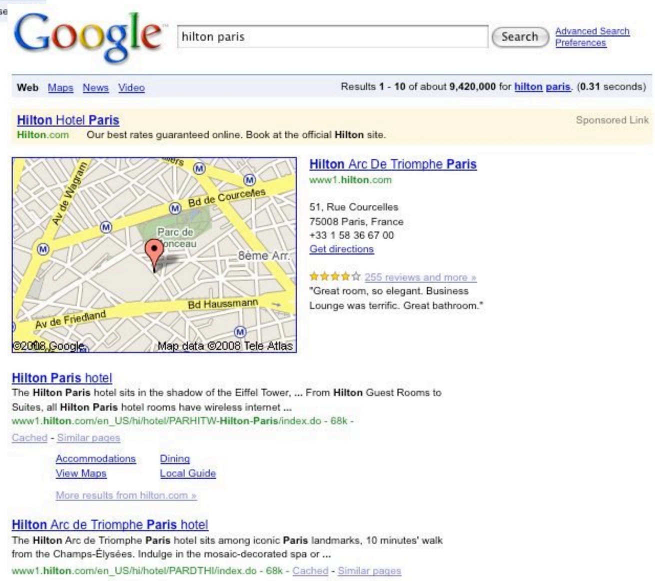


“Personalities” - google.co.uk



paris hilton vs hilton paris - google.com





Cities - google.com

Google | lisbon | Search | 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
Get directions
Remember this location

Lisbon - Wikipedia, the free encyclopedia
Lisbon (Portuguese: Lisboa, IPA: [liʒˈboɐ]) 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 | buchhandlung wien | Suche | 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.baq.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.meine**buchhandlung**.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

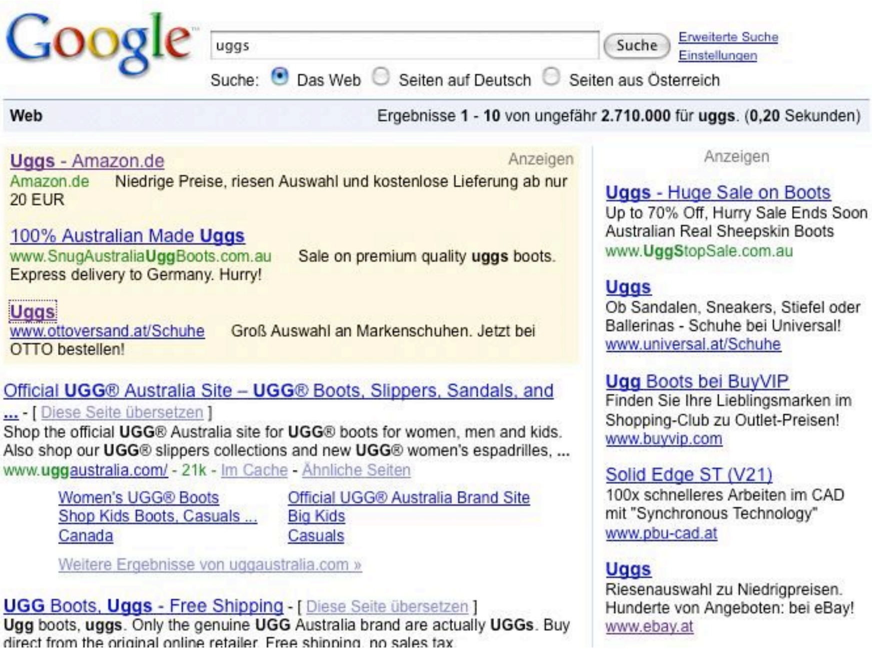
Buchladen
Bücher von A-Z: Große Auswahl, kleine Preise, schnelle Lieferung.
www.bol.at

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

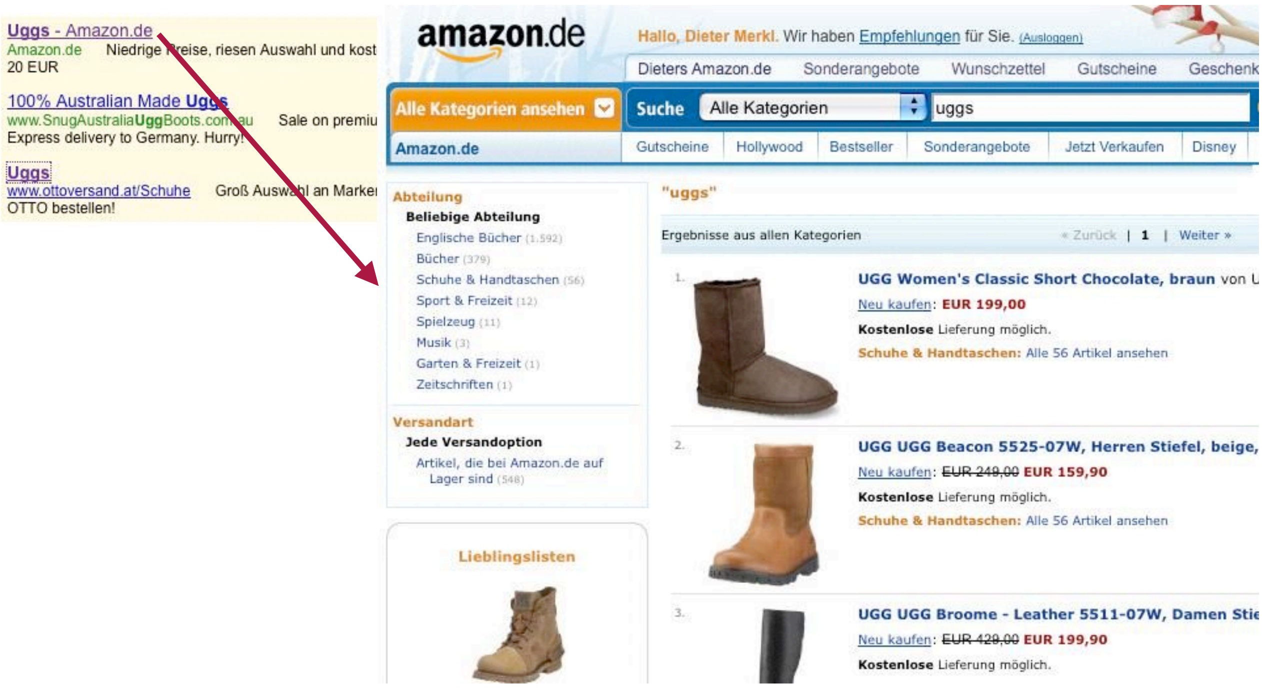
Bücher bei Thalia.at

ec electronic commerce group
Institute of Software Technology and Interactive Systems
TU VIENNA

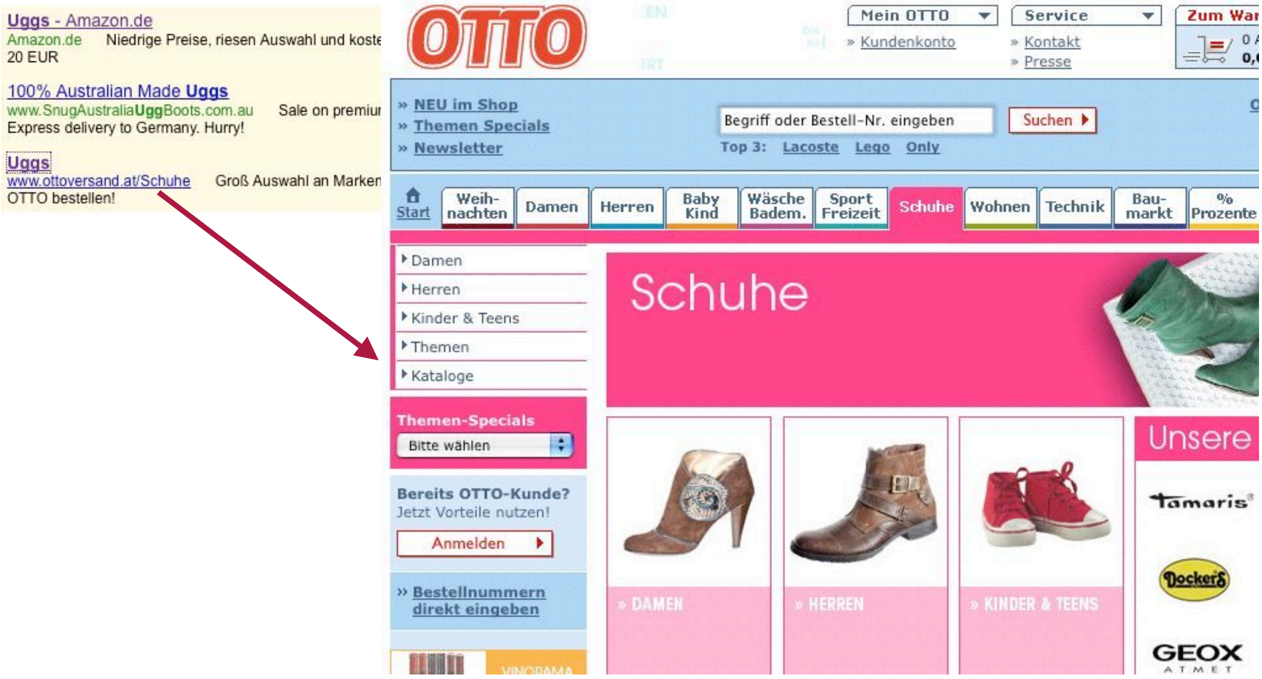
Context transfer - google.at



Context transfer - google.at



No transfer - google.at



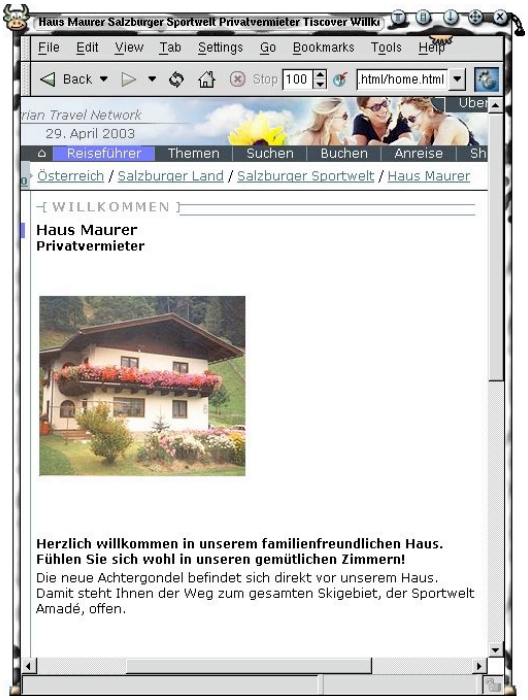
Where to go from here?

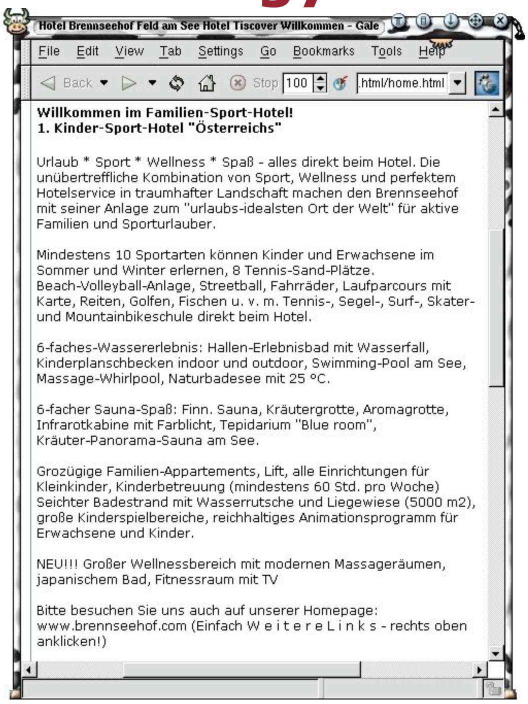
- Text mining
- Concept discovery

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





ec electronic commerce group
Institute of Software Technology and Interactive Systems
TU VIENNA

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

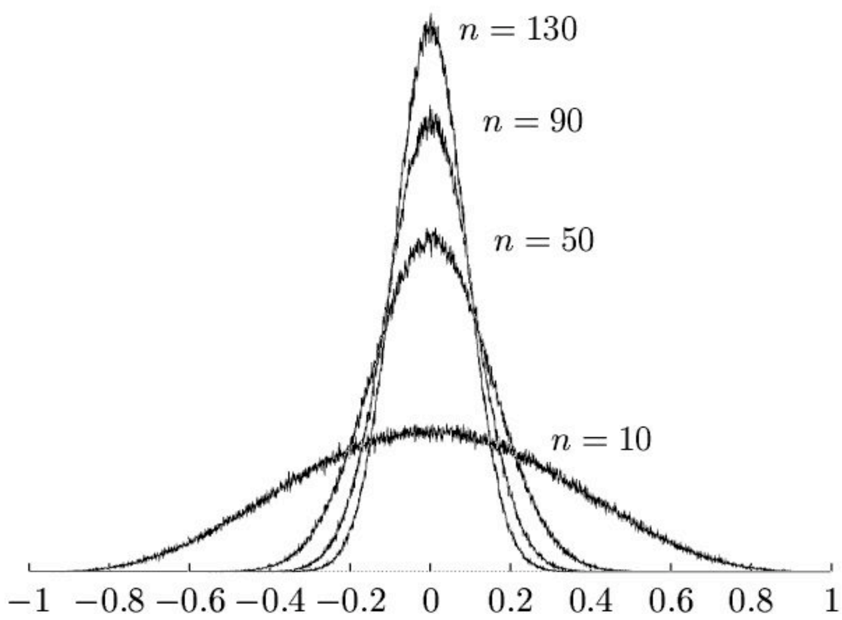
- remove words other than nouns and proper names to avoid primary clustering according to word classes
  - select words starting with capital letter in german texts
  - part-of-speech taggers possible for other languages

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

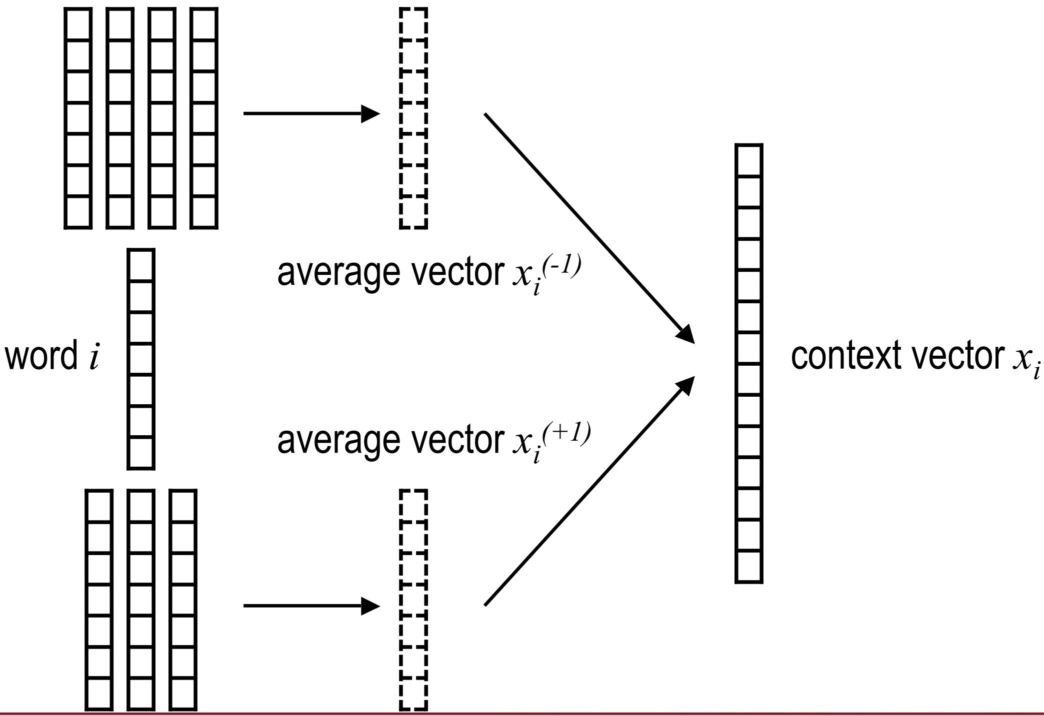
Text mining - Ontology



Text mining - Ontology

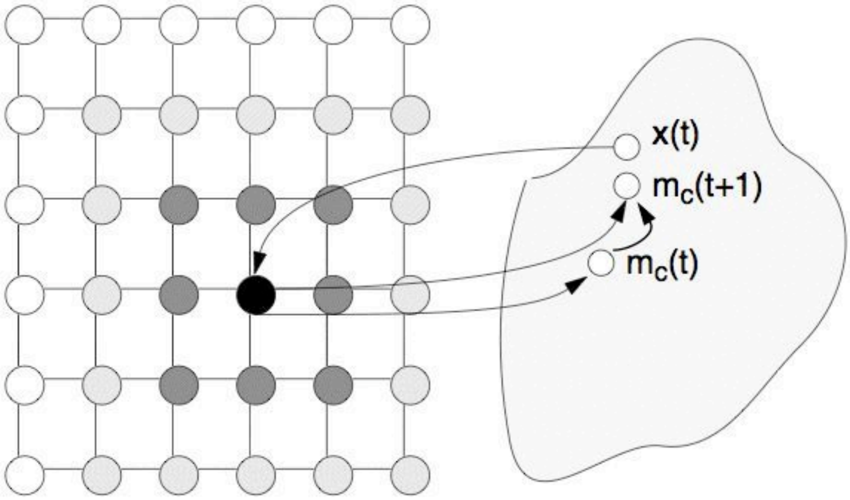
- a list of terms at different displacements is created for each word (e.g. all directly preceding terms at position -1)
- average vectors are calculated => average context
- average context vectors are concatenated to create a vector description of a word determined by its surrounding words
- example: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

Text mining - Ontology

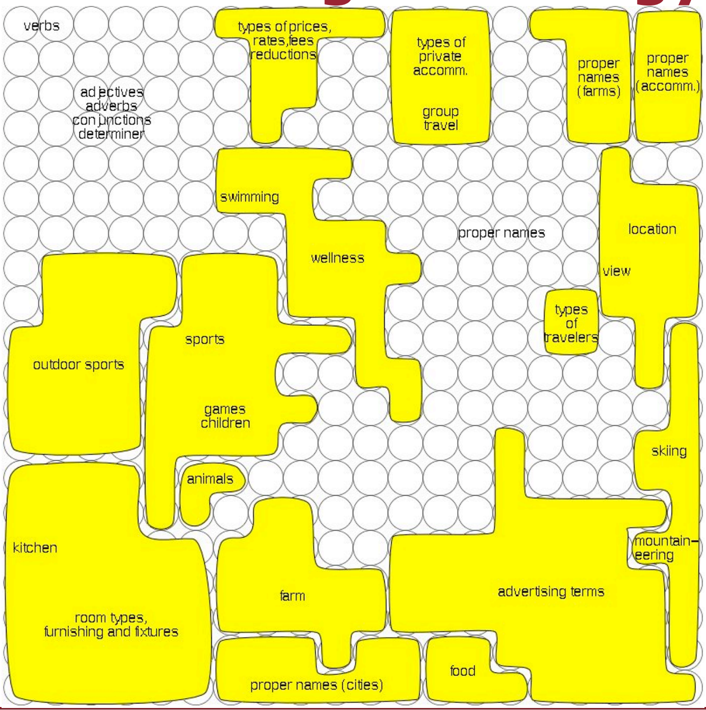


Text mining - Ontology

- Self-organizing map



Text mining - Ontology



Text mining - Ontology

- Detail - lower left corner

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

ec electronic commerce group
Institute of Software Technology and Interactive Systems
TU VIENNA

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

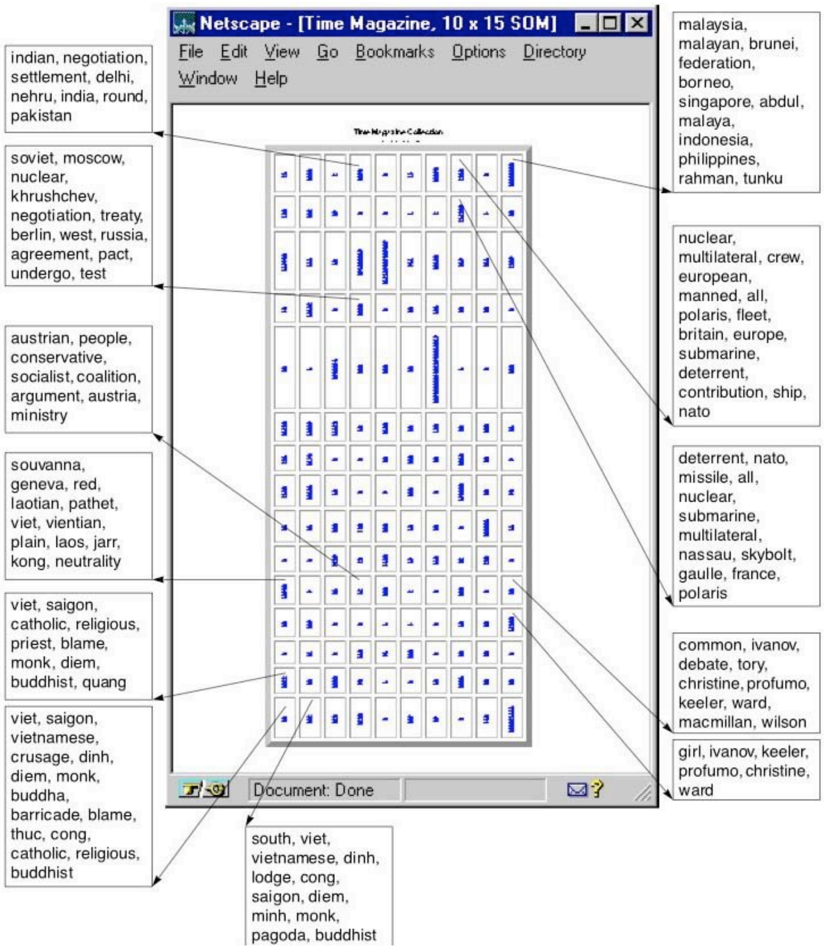
Terms	Federal States								
	Vie	Low. A	Upp. A	St	Bgl	Sbg	Car	Tyr	Vbg
Salzkammergut	0	0	0.8	0.14	0	0.06	0	0	0
Salzkammergutes	0	0	0.76	0.11	0	0.13	0	0	0
Salzkammergutseen	0	0	0.89	0	0	0.11	0	0	0
Arlberg	0	0	0	0	0	0	0	0.11	0.89
Arlberger	0	0	0	0	0	0	0	0.15	0.85
Arlbergs	0	0	0	0	0	0	0	0.02	0.98
Thermenland	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

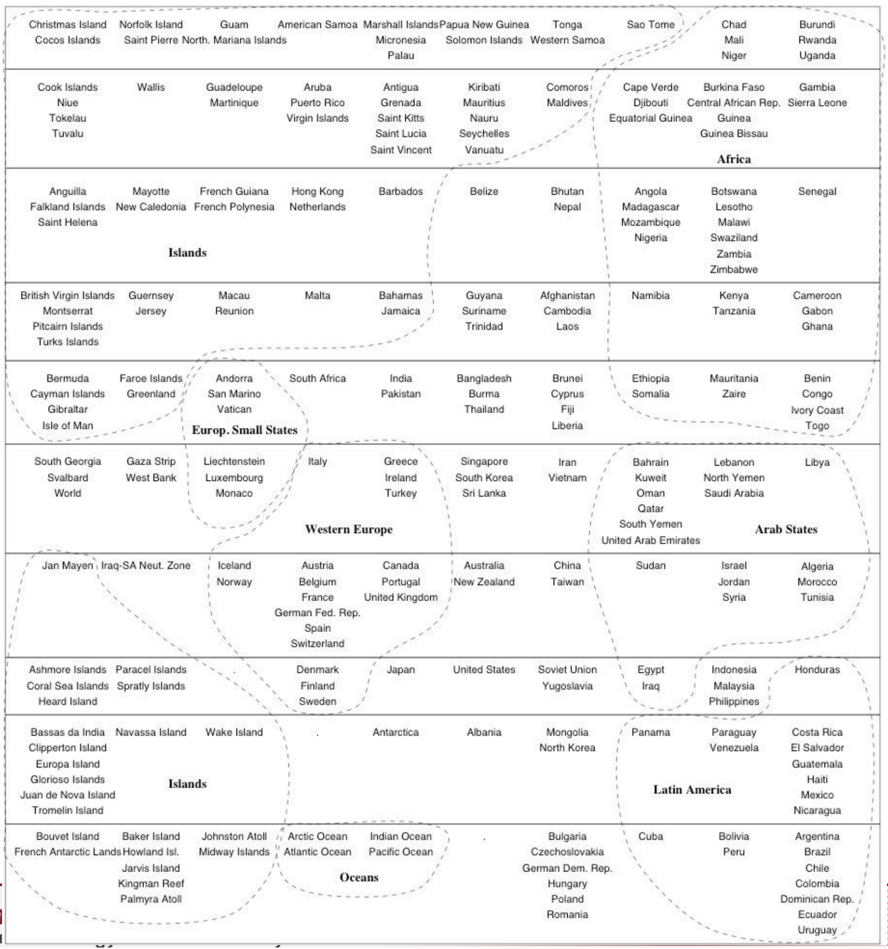
- Goal: Grouping of "similar" documents, i.e. documents covering a "similar" topic

- "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
Document: Done
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
south, viet, vietnamese, dinh, lodge, cong, saigon, diem, minh, monk, pagoda, 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



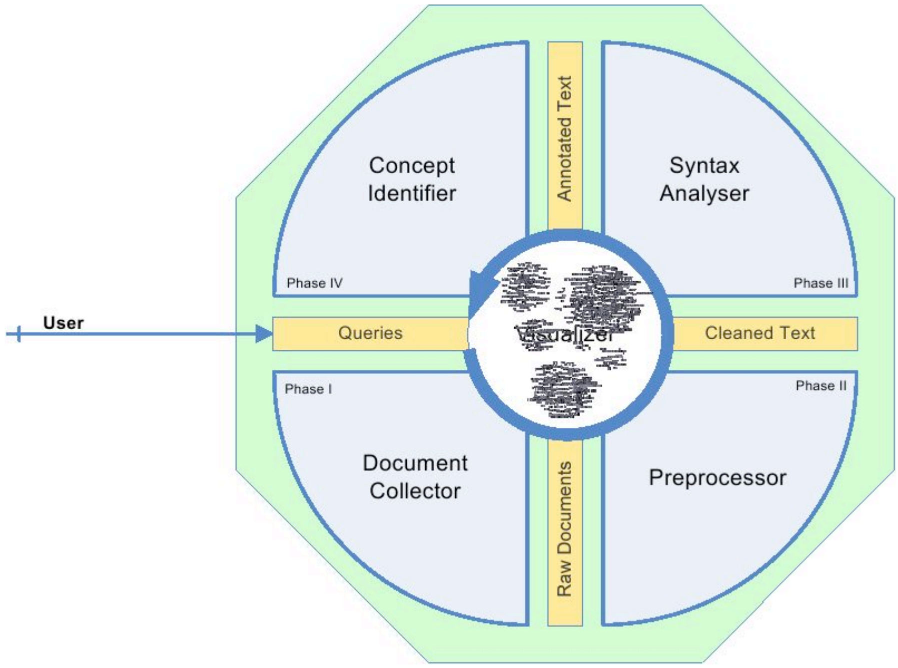
Christmas Island Cocos Islands
Norfolk Island Saint Pierre
Guam North. Mariana Islands
American Samoa
Marshall Islands Micronesia Palau
Papua New Guinea Solomon Islands
Tonga Western Samoa
Sao Tome
Chad Mali Niger
Burundi Rwanda Uganda
Cook Islands Niue Tokelau Tuvalu
Wallis
Guadeloupe Martinique
Aruba 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
Africa
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
Islands
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
Europ. Small States
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 Kuweit Oman Qatar South Yemen United Arab Emirates
Lebanon North Yemen Saudi Arabia
Libya
Western Europe
Arab States
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
Islands
Antarctica
Albania
Mongolia North Korea
Panama
Paraguay Venezuela
Costa Rica El Salvador Guatemala Haiti Mexico Nicaragua
Latin America
Bouvet Island French Antarctic Lands
Baker Island Howland Isl. Jarvis Island Kingman Reef Palmyra Atoll
Johnston Atoll Midway Islands
Arctic Ocean Atlantic Ocean
Indian Ocean Pacific Ocean
Oceans
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

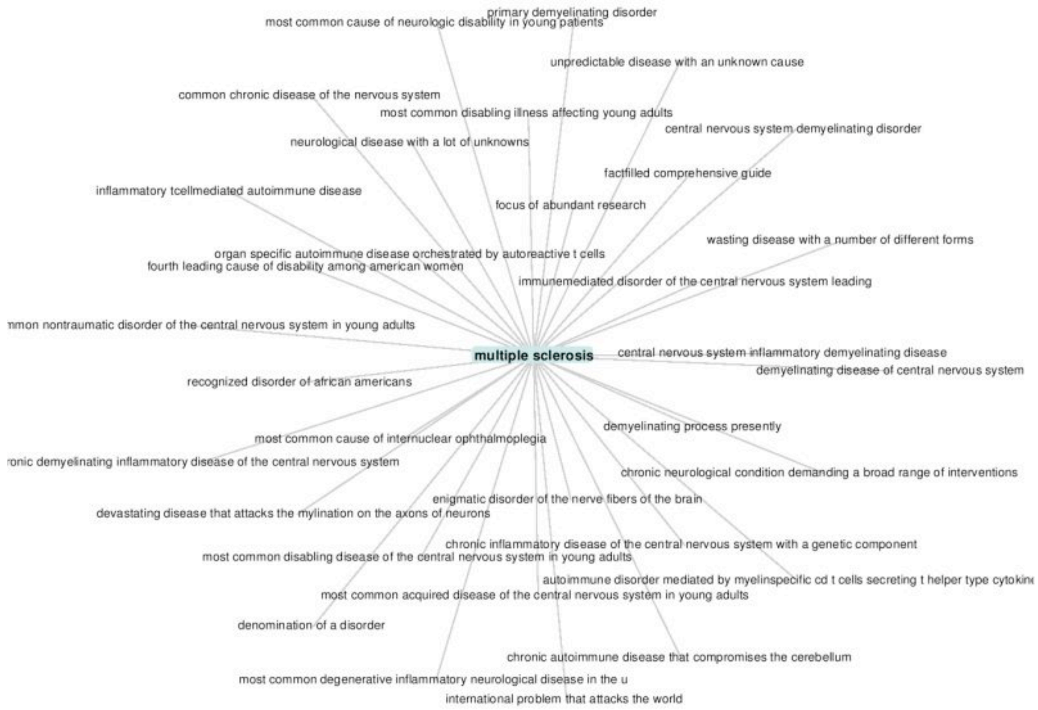
Concept discovery

- 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



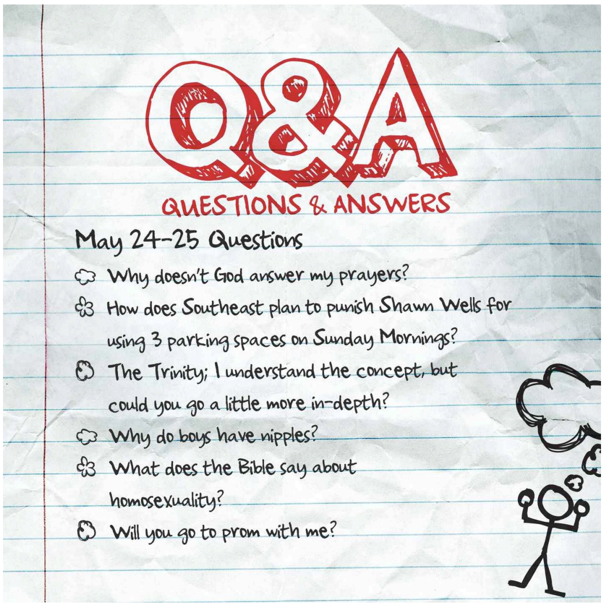
Interesse noch nicht komplett vergangen?

- 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/

ec electronic commerce group
Institute of Software Technology and Interactive Systems
TU VIENNA

Gibt’s Fragen?



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