

# Hinter den Suchmaschinen

Franz Guenther, CIS, LMU München

Arnoud de Kemp, digiprimo

---

- Haben sich Suchmaschinen in den letzten 10 Jahren wesentlich weiterentwickelt?
- Grundsätzliche gelöste und ungelöste Probleme bei der Konstruktion von Suchmaschinen
- Neue Perspektiven für die Benutzerführung und Unterstützung

# 10 Jahre Web-Suchmaschinen

---

- 1997 AltaVista dominiert (300 Millionen Seiten); einige andere Suchmaschinen; Google gegründet
- 1998 Alltheweb startet; Google verhandelt mit AltaVista; Compaq übernimmt Digital
- 1999 Compaq ruiniert AltaVista; technisches Team verlässt AltaVista im Mai 1999
- 2000 Yahoo lizenziert Suche von Google  
Google startet AdWords
- 2001 Google wird populärste Suchmaschine

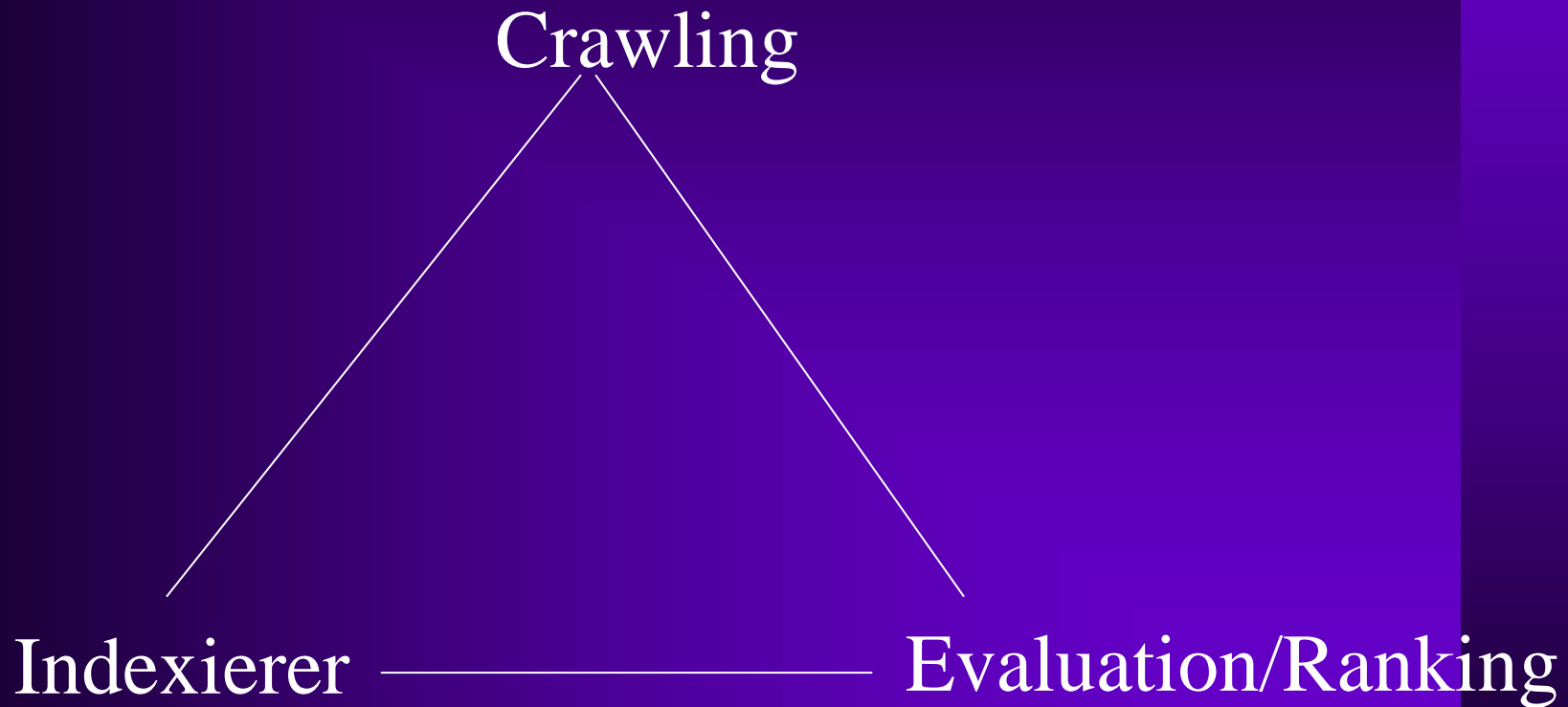
# 10 Jahre Web-Suchmaschinen

---

- 2003 Overture kauft Inktomi, AltaVista AlltheWeb; Yahoo kauft Overture
- 2004 Keine anderen nennenswerten Suchmaschinen; Google erkennt Litfaßsäulen-Effekt der Kombination von Suche und Ad Impressions
- 2007 Alle Suchmaschinen der ersten Generation (inklusive Google) funktionieren weiterhin nach dem gleichen Schema

# Architektur

---



# Query Evaluation

---

- Resultate entstehen durch die Evaluation der Query Terme (fast immer ein logisches „UND“)
- und durch eine statische und eine dynamische Ranking Funktion
- Googles zentraler Beitrag: Einbeziehung von Anchor Text Gewichtung beim Ranking UND bei der Evaluation

# Query Beantwortung

A B C

- Immer noch ca. 2,4 Query Terme per Frage
- Immer noch Listen von Resultaten mit Teasern
- Mehr Flash-ins als früher (erste Flash-in stammen von RealNames im 1997/8)
- Versuch „Antworten“ einzuflashen, wenn das Query „erkannt“ wird („What is the currency of Venezuela?“ oder „two cups in teaspoons“ führt zu der Antwort „two US cups = 96 US teaspoons“)

# Query Typen

---

- Anfragen an eine Suchmaschine können in drei grobe Typen eingeteilt werden:
  - Queries „by description“ (Gelbe Seiten Queries)
  - Queries „by acquaintance“ (Weiße Seiten Queries)
  - Almanac queries
- Somit WWW = (neue Form von) Gelben Seiten

# „Relevanz“ 1: Verschiedene Query Typen

## ■ Die Query x Dokumenten Matrix

Allgemeine Queries			X
Problem Queries	X		
Spezifische Queries		X	
	Inhalt	Format	Referenz

# „Relevanz“ 2

- Man misst oft die Qualität von Suchmaschinen mit den Begriffen „recall“ und „precision“ bezüglich der „relevanten“ Dokumente
- Die Menge der „relevanten“ Dokumente ist aber im allgemeinen schwer zu bestimmen
- Es ist viel leichter Kriterien für irrelevante oder fehlende Ergebnisse zu finden
- 2 Hauptgründe für schlechte Ergebnisse:
  - orthographische Details (Schreibfehler, linguistische Varianten)
  - semantische Details (Synonyme)

# Hauptprobleme bei der Evaluation

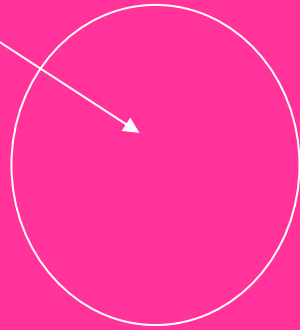
---

- Problem 1: Nur Dokumente die die Queryterme genau in der verwendeten Form enthalten, können überhaupt gefunden werden. Die geringste Variation resultiert (fast immer!) in verschiedenen Resultaten.
- Problem 2: Alle Queries werden hinsichtlich der Evaluation und dem Ranking auf die gleiche Art und Weise behandelt. Weder die Sprache noch die Form des Queries spielt eine Rolle.
- Somit wird in der Resultatpräsentation der „Intention“ hinter den Queries keine Rechnung getragen.

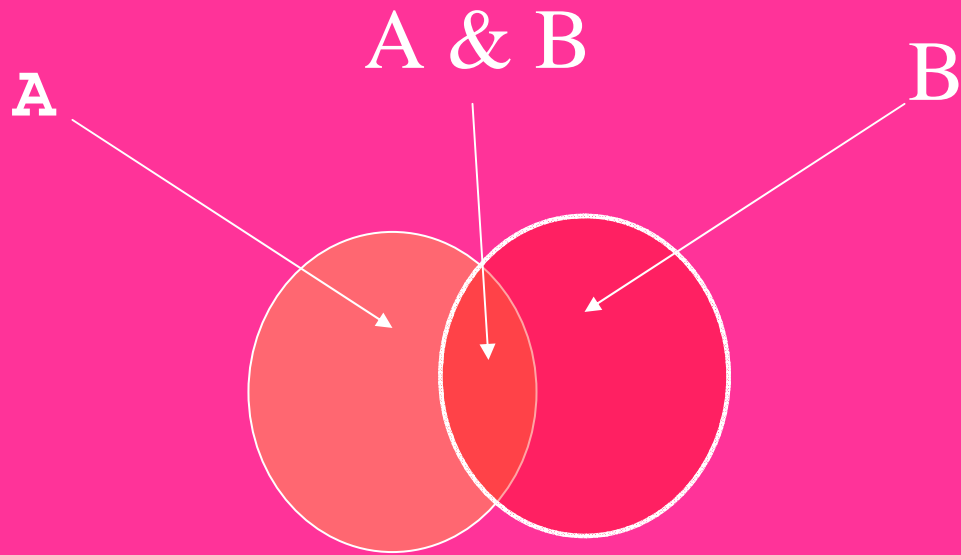
# Evaluation von A B C

---

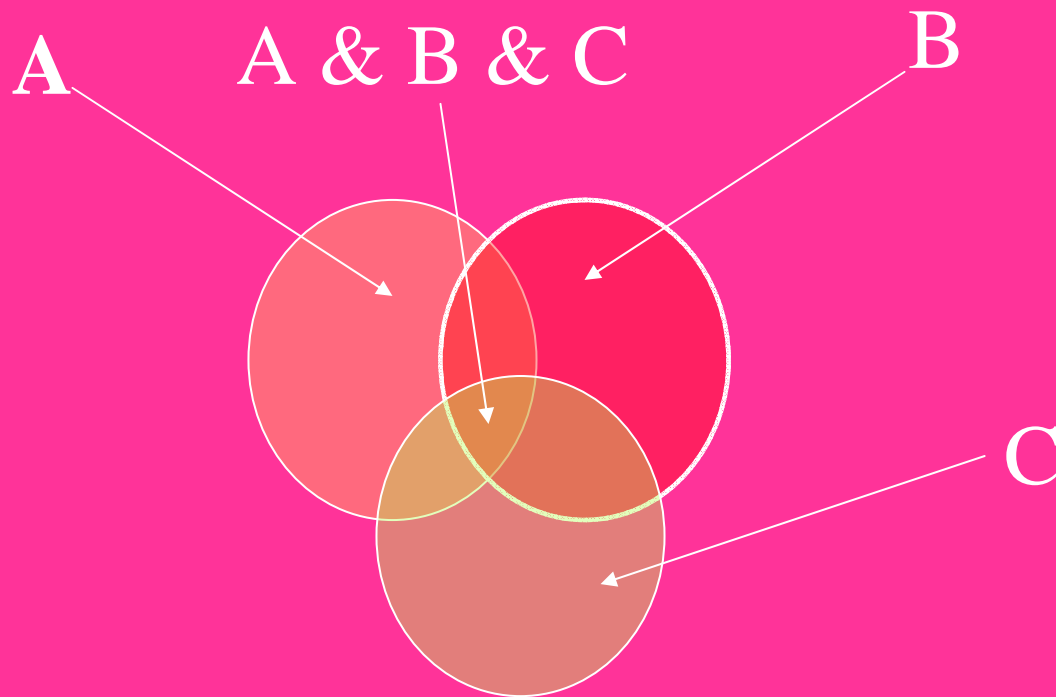
A



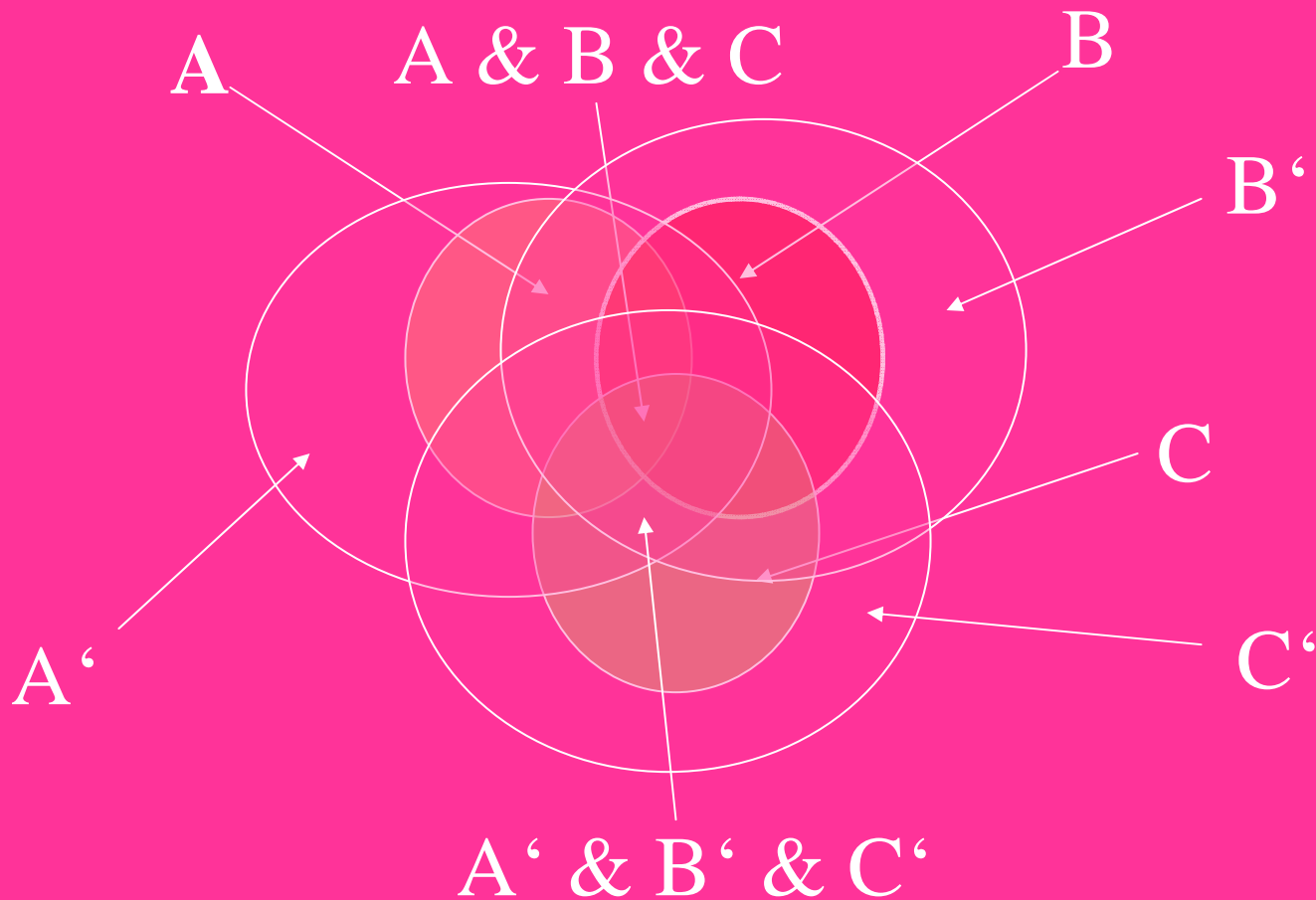
# Evaluation A & B



# Evaluation A & B & C



# A & B & C und die Erweiterung der Evaluation



# Variation in den Ergebnissen

---

- Typisches Beispiel:
- *stellenangebot lungenarzt frankfurt main = 31 Treffer*  
*vs. stellenangebot pneumologe frankfurt main = 56 Treffer*
- Varianten dieser Queries mit Singular/Plural, Maskulin/Feminin oder anderen Schreibformen (z.B. frankurt/main) resultieren in sehr unterschiedlichen Ergebnissen
- Der Benutzer weiß nie, ob die Qualität der Ergebnisse vom Datenbestand abhängt oder von der Formulierung seiner Anfrage

# Ein etwas anderer Ansatz

---

- Der Benutzer soll - so schnell wie nur möglich – sehen, was er überhaupt finden kann
- Das betrifft alle Ebenen der Suche - von Kategorien (sitemaps) bis hin zu den konkreten Termen auf der jeweiligen Webseite
- Diese Vorschlagsgenerierung soll auch orthographisch robust sein: d.h. Fehler jeglicher Art sollten die Vorschlagsgenerierung nicht beeinträchtigen
- Dazu muss man aber in der Lage sein, die „interessanten“ Terme aus den Seiten zu ermitteln (Computerlinguistikproblem)
- Dazu braucht man sehr effiziente Suchalgorithmen, die in der Lage sind, auch aus Millionen von Datensätzen in weniger als 10 ms Antworten zu liefern

# Abhilfe 1: Vorschlagsgenerierung



## < Die Jobsuchmaschine >

Bitte nur Berufs- und/oder Regionsbezeichnungen verwenden

automech		SUCHEN
Automechaniker/in	9100	Herisau (1)
Automechaniker/in	7000	Chur (1)
Automechaniker/in	8400	Winterthur (2)
Automechaniker/in	1950	Sion (CH) (1)
Automechaniker/in	4500	Solothurn (2)
Automechaniker/in	6000	Luzern (1)
Automechaniker/in	9000	St. Gallen (2)
Automechaniker/in	8500	Frauenfeld (1)
Automechaniker/in	4000	Basel (1)
Automechaniker/in	5000	Aarau (8)
Automechaniker/in	3000	Bern (1)
Automechaniker/in	8000	Zürich (2)
Automechaniker/in		Baden (1)
Automechaniker/in		Diverse Standorte (5)
Kraftfahrzeug-Mechaniker/in	96253	Untersiemau (1)
Kraftfahrzeug-Mechaniker/in	91471	Illesheim (1)
Kraftfahrzeug-Mechaniker/in	48308	Senden (1)
Kraftfahrzeug-Mechaniker/in	34289	Zierenberg (1)
Kraftfahrzeug-Mechaniker/in	08141	Reinsdorf (1)
KFZ-Schlosser/in	15837	Baruth (1)
Kraftfahrzeug-Mechaniker/in	82491	Grainau (1)
... tippen Sie einfach weiter um exaktere Ergebnisse zu bekommen ...		

Anzeigen: Job

- Technical Consu
- Senior Sales Ex
- Vertriebsprofi S
- Bürokräft
- IT Architect
- Marketing-Leite
- Presales Consul



Home Hilfe

Francisco, US



Jobs nach Region

Berufe: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Orte: A B C D E F G H I J K L M N O P Q R S T U V W X Z

Berlin Bremen Dresden Düsseldorf Frankfurt M. Hamburg Hannover Köln München Nürnberg Stuttgart »

# Abhilfe 2: Approximative Suche



## < Die Jobsuchmaschine >

Bitte nur Berufs- und/oder Regionsbezeichnungen verwenden

kraftfahrzeugmekan		SUCHEN
Kraftfahrzeug-Mechaniker/in	Amersfoort (1)	
Kraftfahrzeug-Mechaniker/in	Kindersley (3)	
Kraftfahrzeug-Mechaniker/in	Melfort (1)	
Kraftfahrzeug-Mechaniker/in	Odense NV (1)	
Kraftfahrzeug-Mechaniker/in	21698 Harsefeld (1)	
Kraftfahrzeug-Mechaniker/in	94522 Wallersdorf (1)	
Kraftfahrzeug-Mechaniker/in	90613 Großhabersdorf (1)	
Kraftfahrzeug-Mechaniker/in	96253 Untersiemau (1)	
Kraftfahrzeug-Mechaniker/in	91471 Illesheim (1)	
Kraftfahrzeug-Mechaniker/in	48308 Senden (1)	
Kraftfahrzeug-Mechaniker/in	34289 Zierenberg (1)	
Kraftfahrzeug-Mechaniker/in	08141 Reinsdorf (1)	
Kraftfahrzeug-Mechaniker/in	82491 Grainau (1)	
Kraftfahrzeug-Mechaniker/in	85395 Attenkirchen (1)	
Kraftfahrzeug-Mechaniker/in	16341 Panketal (1)	
Kraftfahrzeug-Mechaniker/in	01762 Schmiedeberg (2)	
Kraftfahrzeug-Mechaniker/in	14552 Michendorf (1)	
Kraftfahrzeug-Mechaniker/in	85445 Oberding (1)	
Kraftfahrzeug-Mechaniker/in	82194 Gröbenzell (1)	
Kraftfahrzeug-Mechaniker/in	56412 Heiligenroth (1)	
Kraftfahrzeug-Mechaniker/in	83416 Saaldorf-Surheim (1)	

... tippen Sie einfach weiter um exaktere Ergebnisse zu bekommen ...

Berufe: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Orte: A B C D E F G H I J K L M N O P Q R S T U V W X Z

Berlin Bremen Dresden Düsseldorf Frankfurt M. Hamburg Hannover Köln München Nürnberg Stuttgart »

Anzeigen: Job

Technical Consu  
Senior Sales Ex  
Vertriebsprofi S  
Bürokräft  
IT Architect  
Marketing-Leite  
Presales Consul

VOITH

Home Hilfe

Francisco, US



Platz für  
Ihr Logo!

Jobs nach Region

# Abhilfe 3: Synonym-Erweiterung



## < Die Jobsuchmaschine >

Bitte nur Berufs- und/oder Regionsbezeichnungen verwenden

edv-bera		SUCHEN
EDV-Berater/in	10249	Berlin (1)
EDV-Berater/in	52070	Aachen (1)
IT-Consultant		Niedersachsen (2)
IT-Berater/in		Niedersachsen (1)
IT-Consultant		Luxembourg (3)
IT-Consultant		Baden-Württemberg (2)
IT-Consultant		Nordrhein-Westfalen (2)
IT-Berater/in		Bayern (2)
IT-Consultant		Bayern (2)
IT-Berater/in		Nordrhein-Westfalen (1)
IT-Consultant		verschiedene Standort... (1)
DV-Berater/in	2xxxx	Hamburg (5)
IT-Berater/in	8033x	München (5)
IT-Consultant	10xxx	Berlin (7)
DV-Consultant	10xxx	Berlin (3)
DV-Berater/in	8xxxx	München (9)
IT-Consultant	8xxxx	München (30)
IT-Consultant	2xxxx	Hamburg (15)
IT-Berater/in	20xxx	Hamburg (6)
DV-Consultant	8xxxx	München (9)
DV-Consultant	2xxxx	Hamburg (5)

... tippen Sie einfach weiter um exaktere Ergebnisse zu bekommen ...

Berufe: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Orte: A B C D E F G H I J K L M N O P Q R S T U V W X Z

Anzeigen: Job

Technical Consu  
Senior Sales Ex  
Vertriebsprofi S  
Bürokräft  
IT Architect  
Marketing-Leite  
Presales Consul

VOITH

Home Hilfe

Francisco, US



Platz für  
Ihr Logo!

Jobs nach Region

# Abhilfe 4: Umgebungssuche



## < Die Jobsuchmaschine >

Bitte nur Berufs- und/oder Regionsbezeichnungen verwenden

edv-berater mainz

SUCHEN

### Ergebnisse in **Mainz** und Umgebung

EDV-Berater/in		Mainz (-)
IT-Consultant	6518x	Wiesbaden (2)
DV-Berater/in	65185	Wiesbaden (1)
DV-Consultant	65185	Wiesbaden (1)
IT-Consultant	65843	Sulzbach (Hessen) (3)
DV-Consultant	65843	Sulzbach (Hessen) (3)
DV-Berater/in	65843	Sulzbach (Hessen) (3)
Berater/in	55130	Mainz (1)
SAP-Berater/in	55116	Mainz (2)

### Anzeigen: Job

Technical Consu  
Senior Sales Ex  
Vertriebsprofi S  
Bürokräft  
IT Architect

Marketing-Leiter (m/w)

Presales Consultant (m/w)

Crossroads Europe GmbH

Crossroads Europe GmbH

Schwäbisch Gmünd

Schwäbisch Gmünd

Francisco, US

# Alles auf einmal!



## < Die Jobsuchmaschine >

Bitte nur Berufs- und/oder Regionsbezeichnungen verwenden

metallhilfe passau

SUCHEN

### Ergebnisse in **Passau** und Umgebung

Metallhelfer/in		Passau (-)
Metallhilfskraft	94032	Passau (1)
Metallhilfsarbeiter/in	94104	Tittling (1)

Anzeigen: Job

Technical Consultant

Autonomy Services

München

Senior Sales Executive

Autonomy

Cambridge, UK - San Francisco, US

# Neue Formen der Vorschlags- generierung

---

- Google Suggest vs. Suggest auf den Daten
- Alle bekannten Anwendungen von Suggest lesen nur eine Liste (von links nach rechts) aus; die gezeigten Vorschläge sind in keiner Weise synchronisiert mit den Daten

# Neue Formen der Vorschlags- generierung

---

- Was soll vorgeschlagen werden?
- Wenn die Dokumente schon semi-strukturiert sind (z.B. bei Produktkatalogen oder Verzeichnisstrukturen) kann man die Daten direkt verwenden
- Bei Volltexten ist es um einiges komplizierter: einzelne Wörter sind in der Regel weniger interessant
- Daher Notwendigkeit der automatischen Extraktion von keyword Phrasen aus den Daten

# Neue Formen der Vorschlags- generierung

---

- Kombination von Suche auf Metadaten und im Volltext
- Ein konkretes Beispiel: Aus mehr als 2000 Büchern wurden ca. 12 Millionen Indexterme und Indexphrasen automatisch auf der Basis sehr umfangreicher „lokalen Grammatiken“ und eines sehr großen Lexikons extrahiert.

# Suche im VLB und Volltexten

---

- Gleichzeitige Suche in einem großem Buchkatalog (dem VLB mit mehr als 1,2 Millionen Titel) sowie in den assoziierten Kategorien und Schlagwörtern UND
- im Volltext von mehr als 2000 elektronisch erfassten Büchern
- Das Ranking sieht vor, dass zuerst die Buchtitel angezeigt werden, die auch im Volltext recherchierbar sind, dann die Schlagwörter und schließlich die Indexterme.

# Beispiel 1: Titeltreffer

## Suche im Verzeichnis lieferbarer Bücher und in VTO-Phrasen

Einfach anfangen zu tippen, z.B. *Ildyko vn Kuerthi* oder *Steuer*

Nur im  
VTO-Bestand suchen

Die Suche erfolgt in  
*Volltext-Keyword*.

Bei Titeln gibt es mehr  
sind im Volltext von

<b>Thomas Hecken: Avantgarde und Terrorismus</b>	Titel (233)
<b>Bruce Hoffman: Terrorismus - Der unerklärte Krieg</b>	Titel (232)
J.A. Corlett: Terrorism	Titel (232)
J. Angelo Corlett: Terrorism	Titel (231)
Volker Pfeifer: Terrorismus	Titel (230)
Thomas Hilker: Terrorismus	Titel (230)
Sascha B Storck: Terrorismus	Titel (230)
Peter Waldmann: Terrorismus	Titel (230)
Kai Hirschmann: Terrorismus	Titel (230)
Jeffrey I Ross: Political Terrorism	Titel (230)
Berndt G Thamm: Terrorismus	Titel (230)
Andrew Silke: Suicide Terrorism	Titel (230)
Rolf Tophoven: GSG 9 - Command against Terrorism	Titel (229)
Michael Gold-Biss: The Discourse on Terrorism	Titel (229)
Mark A Gabriel: Islam und Terrorismus	Titel (229)
Mario Petri: Terrorismus und Staat	Titel (229)
Haig Khatchadourian: The Morality of Terrorism	Titel (229)
Fernando Reinares: Terrorismus Global	Titel (229)
Doron Zimmermann: The Transformation of Terrorism	Titel (229)
Charles Townshend: Terrorismus	Titel (229)

# Beispiel 2: Indextermtreffer

## Suche im Verzeichnis lieferbarer Bücher und in VTO-Phrasen

Einfach anfangen zu tippen, z.B. *Ildyko vn Kuerthi* oder *Steuer*

Nur im  
VTO-Bestand suchen

Die Suche erfolgt in  
*Volltext-Keyword.*

Bei Titeln gibt es me  
sind im Volltext von

macht|

Wolfgang Kuzmits/Katharina Machtinger: Bildführer - Schloss Esterházy - Eisenstadt	Titel (225)
Matthias Braun: Kulturinsel und Machtinstrument	Titel (225)
Dagmar Bez/Felix Osterheider: Der Zölibat - ein Machtinstrument der katholischen Kirche?	Titel (223)
Annette van Edig: Die Nutzung internationaler Wasserressourcen: Rechtsanspruch oder...	Titel (223)
Claudia Y Ludwig: Die nationalpolitische Bedeutung der Ostsiedlung in der Weimarer Repu	Schlagwort (221)
Machtinstanz	Inhaltskeyword (218)
Machtinhaber	Inhaltskeyword (218)
Machtinstinkt	Inhaltskeyword (217)
Machtillusion	Inhaltskeyword (217)
zentrale Machtinstanz	Inhaltskeyword (216)
neuer Machtinhaber	Inhaltskeyword (216)
formeller Machtinhaber	Inhaltskeyword (216)
Machtinteresse	Inhaltskeyword (216)
Machtinstrument	Inhaltskeyword (216)
Machtinsignien	Inhaltskeyword (216)
Machtindikator	Inhaltskeyword (216)
wirksames Machtinteresse	Inhaltskeyword (215)
NS-Machtinszenierung	Inhaltskeyword (215)
Machtinstitution G-BA	Inhaltskeyword (215)
Kampf um Machtinteressen	Inhaltskeyword (215)

Powered by ex@rbyte

# Demolinks

---

- [jobanova.de](http://jobanova.de)
- [rollo.cis.uni-muenchen.de/buchsuche](http://rollo.cis.uni-muenchen.de/buchsuche)
- [rollo.cis.uni-muenchen.de/dtv](http://rollo.cis.uni-muenchen.de/dtv)
- [exorbyte.com](http://exorbyte.com)

# Erweiterungen

---

- Massive Anreicherung der verwendeten Kategorisierung und Schlagwörter und der extrahierten Indexterme durch einen umfangreichen Thesaurus
- Clustering von Indextermen nach verschiedenen Kriterien
- Gewichtung der Indexterme mit verschiedenen Methoden