

LSO ist nur ein Hype ?

Wie berechnet man Semantik ? Gestern – Heute – Morgen

Auf der Suche nach dem semantischen Raum

Glaskugel

Ist noch etwas Zeit übrig ?
→ Themenrelevante Links

Wer spricht da eigentlich ?

Semager

Semantic Business – semantische Datenbanken und Dienstleistungen.
(Keyword-Datenbanken, Konzept-Extraktionen, Kategorisierung von Webseiten, Texten oder einzelnen Wörtern, ...)

www.semager.de = Showcase


Referenzen:

- T-Online
- Yahoo Partner
- Kelkoo Partner
- Sistrix
- Mediadonis
- Domainparking Sites

Matthias Schneider

- Geschäftsführer Semager
- Referent auf der SES, SEMSEO, SEOCampixx, ...
- Seminarleiter in verschiedenen Weiterbildungskursen und Handelskammern (SEO)
- Gastautor in Fachzeitschriften

LSO – Latent Semantische Optimierung



Web · [Preisvergleich](#) · [Verwandte W](#)

Einfache Berechnung

Verwandte Wörter [?] **Abgehende Wortbeziehungen [?]**

1 - 25 von 1.481 **1 - 25 von 176**

- 101% festnetz
- 99% t home
- 96% flatrate
- 96% handy
- 93% telefonieren
- 93% anbieter
- 93% dsl anbieter
- 92% telefonanschluss
- 91% entertain
- 88% dsl anschlüsse
- 87% adsl
- 86% privatkunden
- 86% breitband
- 86% dsl geschwindigkeit
- 86% bereitstellungspreis
- 85% isdn
- 85% verfügbarkeit
- 84% tarife
- 84% meldungen
- 84% surf
- 84% versatel
- 83% speedtest
- 83% mobilfunk
- 83% telekommunikation
- 82% anschluss

61% internet

55% deutschland

55% deutsche

Search Terms:

Language:

Concentration Energy (CEN):

Max Cascading Level (MCL):

telekom (de): 87 outgoing / treshold: 478

dsl (de): 107 outgoing / treshold: 1220

Fired words total: 3673

Seen words total: 3833

Word	How many times fired	Treshold
tarife	50	387
festnetz	38	255
monat	36	735
telefonieren	35	239
mobilfunk	34	338
surfen	32	338
t mobile	31	372
call	29	414
flatrate	29	184
sms	28	634
mobil	25	913
kabel	25	741
voip	25	282
iphone	24	992
anbieter	23	1802
handy	23	1756
tarif	22	267

Neuronale Berechnung

LSO ist nur ein Hype ?

Fakt:

Google ergänzt normale Websuchen um Synonyme

LSO ist nur ein Hype ?

Problem

Es werden Internetseiten gefunden in denen die eingegebenen Suchbegriffe vorkommen (Volltextsuche). Seiten die Inhaltlich dem gesuchten entsprechen, aber die Suchbegriffe nicht enthalten, werden auch nicht angezeigt.

Lösung

Es werden Internetseiten gefunden, in denen nicht unbedingt die Suchbegriffe selbst vorkommen müssen, aber doch deren semantische, sprich inhaltliche Schnittpunkte.

Derzeit bleibt einer Suchmaschine nichts anderes übrig als sich die Semantik mit Methoden der Computerlinguistik zu errechnen. Dazu gibt es generell drei verschiedene Wege:

- 1) Der natürliche Sprachprozess
- 2) Das verstehen des Web-Kontextes
- 3) Das verstehen der Nutzerabsicht

Semantische Suche

1) Der natürliche Sprachprozess

Stichwort: NLP (natural language processing), Computerlinguistik

- Korrektur von Tipp- und Rechtschreibfehlern
- Prüfung auf grammatische Richtigkeit
- Automatische Übersetzung.
- Verschlagwortung von Literatur
- Anfertigung von Registern und Inhaltsverzeichnissen
- Herstellung von Zusammenfassungen und Abstracts.
- Unterstützung von Autoren beim Verfassen von Texten

Semantische Suche

2) Das verstehen des Web-Kontextes

- Homonyme - gleiches Wort kann je nach Kontext andere Bedeutung haben.
- Auflösung syntaktischer Mehrdeutigkeiten - ein Satz lässt sich auf mehrere Arten deuten.

Beispiel:

„Peter sah Maria mit dem Fernglas“

Hat Peter Maria gesehen hat, die

- a) ein Fernglas in der Hand hielt, oder hat
- b) Peter Maria mit Hilfe eines Fernglases gesehen ?

Semantische Suche

3) Das verstehen der Nutzerabsicht

Ein Suchender gibt das Wort „Kamera Canon EOS 450D“ ein.
Interessiert er sich nun für einen Preisvergleich oder einen Testbericht?

- Navigationsorientiert – der Nutzer sucht den Hersteller
- Informationsorientiert – der Nutzer sucht Testberichte
- Transaktionsorientiert – der Nutzer sucht Anbieter
- Evtl. auch weitere wie z.B. Ressourcen (Downloads) oder Media (Videos, Bilder)

→ Erkenntnis nutzen von Nutzerprofile, Standort, Suchhistorie, Klickhistorie

LSO ist nur ein Hype ?

Society for Experimental Mechanics (SEM) - [[Diese Seite übersetzen](#)]

The development and application of engineering measurements and test methods to the determination of materials and system behavior.

www.sem.org/ - [Im Cache](#) - [Ähnlich](#)

SEM-Tools: Konkurrenzanalyse, Brandprotection, Marktbeobachtung ...

SEM, also die Buchung bezahlter Anzeigen dort, wo der Kunde sie wahrnimmt, hat in den letzten Jahren ein beispielloses Wachstum erfahren. Das SEM-Modul der ...

<https://tools.sistrix.de/sem> - [Ähnlich](#)

Scanning electron microscope - Wikipedia, the free encyclopedia - [[Diese Seite übersetzen](#)]

The scanning electron microscope (SEM) is a type of electron microscope that In a SEM, as in scanning probe microscopy, magnification results from the ...

en.wikipedia.org/.../Scanning_electron_microscope - [Im Cache](#) - [Ähnlich](#)

Search engine marketing - Wikipedia, the free encyclopedia - [[Diese Seite übersetzen](#)]

Search engine marketing, or SEM, is a form of Internet marketing that seeks to promote websites by increasing their visibility in search engine result pages ...

en.wikipedia.org/wiki/Search_engine_marketing - [Im Cache](#) - [Ähnlich](#)

SEM = Society for Experimental Mechanics
= Scanning electron microscope
= Search engine marketing

LSO ist nur ein Hype ?

Google [Erweit](#)

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web [+ Optionen anzeigen...](#) Ergebnisse 1

[Lehrstellen Hannover](#)
markt.de/Jobs/Hannover Suchen Sie nach einem Job in **Hannover**? Jetzt Traumjob finden

[Lehrstellen hannover, Ausbildungsplätze hannover, Lehrstelle ...](#)
Lehrstellen hannover, Ausbildungsplätze hannover, Ausbildungsplatz hannover, **Lehrstelle hannover**, Lehrstelle, Ausbildungsplätze, Ausbildungsplatz, ...
www.backinjob.de/lehrstellen/lehrstellen-hannover - vor 11 Stunden gefunden -
[Im Cache](#) - [Ähnlich](#) - [🗨](#) [📄](#) [✕](#)

[Lehrstellen Hannover, Ausbildungsplatz Hannover - meinestadt.de](#)
Ausbildungsplätze und **Lehrstellen** für **Hannover** und Umgebung findest Du im **Lehrstellenmarkt** von meinestadt.de - plus weitere nützliche Infos zu Ausbildung ...
www.meinestadt.de/hannover/lehrstellen - [Im Cache](#) - [Ähnlich](#) - [🗨](#) [📄](#) [✕](#)

Lehrstellen = Lehrstellenmarkt

LSO ist nur ein Hype ?




Größere Stellensuchmaschinen und Jobbörsen - **Stellensuche Berlin**

Übersicht zu größeren Stellensuchmaschinen, Jobbörsen und Metasuchmaschinen die geeignet für die Jobsuche in **Berlin** sind.

www.stellensuche-berlin.de/suchm.htm - [Im Cache](#) - [Ähnlich](#) -   

Jobbörse Berlin Jobsuche Berlin **Stellensuche Berlin**

Auf Jobsuche **Berlin**? Unsere Jobbörse **Berlin** mit 881.055 aktuellen Stellen ist Ihre perfekte Jobsuchmaschine! So macht Jobsuche Spaß!

www.kimeta.de/Jobsuche_Berlin_.aspx - [Im Cache](#) - [Ähnlich](#) -   

Stellenangebote **Berlin**, Stellenmarkt **Berlin**, Jobbörse **Berlin**

Stellenangebote **Berlin**, Stellenmarkt **Berlin**, Stellenangebot **Berlin**, Jobbörse **Berlin**.

www.backinjob.de/stellenmarkt-Berlin - [Im Cache](#) - [Ähnlich](#) -   

Stellengesuche: berlin - Stellenmarkt.de

Stellengesuche: berlin - Stellenmarkt.de Ihr Internet Stellenmarkt, über 33.000 Stellenangebote, 20.000 Bewerber. Einer der führenden Online ...

www.stellenmarkt.de/stellengesuche/berlin - [Im Cache](#) - [Ähnlich](#) -   

Stellensuche = Stellengesuche

LSO ist nur ein Hype ?

Google [Erweiterte Suche](#)
[Einstellungen](#)
Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web

[Stellenanzeigen Berlin](#)

www.monster.de/Berlin Jobs in der Hauptstadt **Berlin**. Jetzt sofort online bewerben!

[Jobs in Berlin](#)

FAZjob.net für Führungskräfte & Spezialisten: Top-Stellenangebote auf FAZjob.NET

[Stellenangebote Berlin](#), [Stellenmarkt Berlin](#), [Jobbörse Berlin](#), [Jobs ...](#)

[Stellenangebote Berlin](#), [Stellenangebot Berlin](#), [Stellenanzeigen Berlin](#), [Jobbörse Berlin](#), [Stellenmarkt Berlin](#), [Jobs Berlin](#).

www.backinjob.de/stellenangebote-Berlin - 49k - [Im Cache](#) - [Ähnliche Seiten](#)

[Jobs Berlin](#), [Stellenangebote Berlin](#) - meinestadt.de

Stellenangebote in **Berlin**: Der regionale Stellenmarkt **Berlin** bei meinestadt.de bietet übersichtlich freie Stellen und Jobs in **Berlin** und Umgebung.

jobs.meinestadt.de/berlin - 81k - [Im Cache](#) - [Ähnliche Seiten](#)

Stellenanzeigen = Stellenangebote

LSO ist nur ein Hype ?

 [Erweiterte Su...](#)

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web [+ Optionen anzeigen...](#) Ergebnisse 1 - 10 von ungefähr 11

[Katholische Kirche in Neustadt a.Rbge./Home](#)
St. Peter und Paul in **Neustadt** am Rbge. Herzlich Willkommen auf der Homepage der Katholischen **Kirchengemeinde**.
www.katholische-kirche-neustadt.de/ - [Im Cache](#) - [Ähnlich](#) -   

[Evangelische Kirchengemeinde Neustadt](#)
Internetpräsenz der **Evangelischen Kirchengemeinde** Waiblingen-Neustadt.
www.elkw.de/gemeinden/neustadt - [Im Cache](#) -   

[Evangelische Kirchengemeinde Neustadt](#)
Etwas über 2200 Gemeindeglieder im Stadtbereich **Neustadt**, den Ortsteilen ... Bubenbach und Schollach bilden die **Evangelische Kirchengemeinde Neustadt**. ...
www.ekineu.de/ - [Im Cache](#) - [Ähnlich](#) -   

LSO ist nur ein Hype ?

Google [Erweiterte Su...](#)

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web [+ Optionen anzeigen...](#) Ergebnisse 1 - 10 von ungefähr 11

[Katholische Kirche in Neustadt a.Rbge./Home](#)
St. Peter und Paul in Neustadt am Rbge. Herzlich Willkommen auf der Homepage der Katholischen **Kirchengemeinde**.
www.katholische-kirche-neustadt.de/ - [Im Cache](#) - [Ähnlich](#) - [🗨](#) [📄](#) [🗕](#)

[Evangelische Kirchengemeinde Neustadt](#)
Internetpräsenz der Evangelischen Kirchengemeinde Waiblingen-Neustadt.
www.elkw.de/gemeinden/neustadt - [Im Cache](#) - [🗨](#) [📄](#) [🗕](#)

[Evangelische Kirchengemeinde Neustadt](#)
Etwas über 2200 Gemeindeglieder im Stadtbereich Neustadt, den Ortsteilen ... Bubenbach und Schollach bilden die Evangelische Kirchengemeinde Neustadt. ...
www.ekineu.de/ - [Im Cache](#) - [Ähnlich](#) - [🗨](#) [📄](#) [🗕](#)

- Sie sehen das nicht „ökumenisch“ genug, Google schon
- Rom lässt sich den Alleinvertretungsanspruch etwas kosten
- Sie haben beim Verwendungszweck für den Kirchenbeitrag „Suchmaschinenoptimierung“ angekreuzt

LSO ist nur ein Hype ?

Google ergänzt/ändert Suchbegriffe

Stellenanzeigen Berlin

Webagentur Berlin

Webseiten Design

→ Stellenangebote Berlin

→ Werbeagentur Berlin

→ Webdesign

song words

what state has the highest murder rate

himalayan kitten breeder

→ „words“ wurde ergänzt durch „lyrics“

→ „homicide“ wurde ergänzt für „murder“

→ "cat breeder" ist das gleiche wie „kitten breeder“

Kontextual:

dura ace track bb axle njs

software update on bb color id

bb cream dark

southeastern usa bb fitness & figure

→ "bb" here means "bottom bracket".

→ „bb“ steht für „blackberry“

→ hier steht „bb“ für „blemish balm“

→ „bb“ steht hier für „bodybuilding“

arm reduction oder arms reduction

→ keine Wortstammreduzierung

Quellen:

<http://googleblog.blogspot.com/2010/01/helping-computers-understand-language.html>

LSO ist nur ein Hype ?

Google ergänzt/ändert Suchbegriffe

Google Patent 7,409,383

Methode, um Synonyme oder anderen Ersatz-Klauseln zu bestimmen.

Für jede Suchanfrage wird eine Vielzahl von Pseudo-Suchanfragen bestimmt, jede Pseudo-Suchabfrage wird abgeleitet von Suchanfragen bei denen ein Phrase ausgetauscht wurde.

Ein potenzielles Synonym ist ein Begriff, der

a) innerhalb einer benutzerdefinierten Abfrage an die Stelle in einer Suchanfrage verwendet wurde

b) und im Kontext einer Pseudo-Suchanfrage auftaucht.

Quellen:

<http://arnoldit.com/wordpress/2009/12/24/google-nails-patent-for-query-synonyms-in-query-context/>

<http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,636,714.PN.&OS=pn/7,636,714&RS=PN/7,636,714>

LSO ist nur ein Hype ?

Google squared (ver)sucht semantisch

Google squared labs

mozart birthday					
	Item Name	Image	Description	Death	Date Of Birth
<input type="checkbox"/>	Wolfgang Amadeus Mozart		Wolfgang Amadeus Mozart (German: [ˈvɔlfɡaŋ amaˈdeʊs ˈmoːtʰsart], full baptismal name Johannes Chrysostomus Wolfgangus Theophilus Mozart (27 January 1756 – 5 ...	Vienna, Austria	27 January 1756
<input type="checkbox"/>	Wolfgang Amadeus Mozart Biography		Wolfgang Amadeus Mozart . (1756 - 1791) Probably the greatest genius in Western musical history, Wolfgang Amadeus Mozart was born in Salzburg, Austria, Jan.	Dec 5, 1791	1 possible value
<input type="checkbox"/>	Franz Xaver Wolfgang Mozart		Franz Xaver Wolfgang Mozart (26 July 1791 – 29 July 1844), also known as F. X. Mozart, W. A. Mozart Son, or Wolfgang Amadeus Mozart, Jr., was the youngest ...	29-Jul-1844	2 possible values
<input type="checkbox"/>	Anna Maria		Maria Anna Mozart, or "Nannerl" as family called her, is the older sister of the ... Maria Anna Nannerl Mozart, Wikimedia Commons Maria Anna Nannerl Mozart ...	2 possible values	4 possible values
<input type="button" value="Add items"/>		<input type="button" value="Add"/>	or Add next 10 items		

LSO ist nur ein Hype ?

Google Branchencenter:

Bei Auswahl der Kategorie, wird die Suche um Synonyme zu dieser Kategorie ergänzt. Webseiten werden mit Synonymen Suchbegriffen gefunden, die nicht auf der Webseite oder in Links zu diesen zu finden sind.

Leider keine Beweisnennung möglich, da Kundenprojekt – jedoch Tatsache.

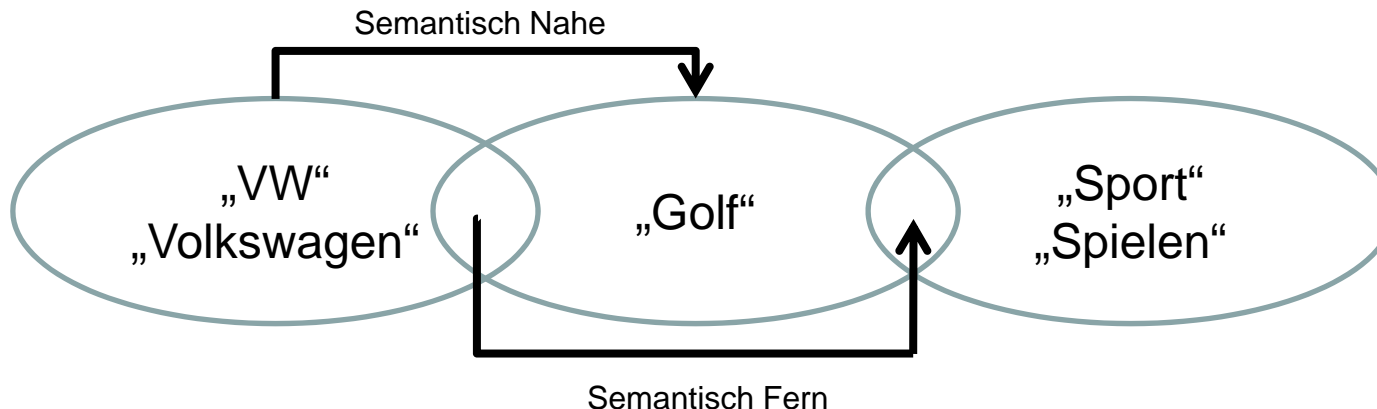
Wie berechnet man Semantik ? Gestern – Heute – Morgen

LSO – Latent Semantische Optimierung

„semantischen Nahe“ $\leftarrow \rightarrow$ „semantisch Fern“

In Webseiten bei denen es um „Golf“ geht, wird auch oft „VW“ und „Volkswagen“ genannt. Webseiten in denen zwar „Golf“ genannt wird, aber in anderem Zusammenhang (stattdessen mit „Sport“ und „Spielen“), sind *semantisch Fern* zu diesen. *Semantisch Nahe* sagt man, wenn Webseiten zwar „VW“ und „Volkswagen“ haben, aber eben nicht „Golf“.

- + Semantisch Nähe
- Längere Such- und Analysezeiten
- Eben doch nur Latent (und deswegen auch nicht *Synonym*)



LSI \leftarrow Scott Deerwester, 1990

LSO – Latent Semantische Optimierung

Content DNA

The closer the content of your webpage matches the ContentDNA, the higher the search engine concerned will rank your webpage on the content score for the specific search term.

Gravitationszentrum eines Themenclusters = ContentDNA

ABER: Common Neurolinguistic Map

es macht einen Unterschied was man von Schnee hält, ob man nun in der Arktis oder in der Karibik wohnt (→ kulturell und geographischer Sprach- und Meinungsraum)

Berechnung von Semantik

Phonetik

Hat eigentlich nichts mit Semantik zu tun, sollte aber der Vollständigkeit halber mal genannt sein: berechnet die Aussprache eines Wortes:

TABLE 1: SOUNDEX and Double Metaphone Encoding

Word	SOUNDEX	Primary Metaphone	Secondary Metaphone
Smith	S530	SM0	XMT
Smythe	S530	SM0	XMT
phone	P500	FN	no code
pony	P500	PN	no code
George	G620	JRJ	KRK
garage	G620	KRJ	KRK
benign	B525	PNN	PNKN
benignant	B525	PNNNT	PNKNNT
poignant	P255	PNNT	PKNNT

Interessant in Verwendung mit Levenshtein-Distanz, um z.B. Falschschreibweisen zu finden.

Soundex ← Robert Russel, 1918

Metaphone ← Lawrence Philips, 1990

Berechnung von Semantik

Thesaurus

Anbindung/Import einer bestehenden Thesaurus Datenbank und Vergleich der Suchanfragen mit dieser

- Lexikalisch
 - Wordnet , Germanet , Wortschatz Uni Leipzig
- GPL
 - Openthesaurus
 - Wikipedias Wictionary
- Fachthesauri:
 - Standardthesaurus Wirtschaft (STW)
 - Medizin
- Multilinguale Thesauri
 - UNESCO
- Kostenpflichtige
 - Dornseiff

wictionary.org:

bugsieren

bugsieren (Deutsch) [Bearbeiten]

Verb [Bearbeiten]

Silbentrennung:

bug sie ren, Präteritum: bug sier te, Partizip II: bug siert

Aussprache:

IPA: [buk'si:ren], Präteritum: [buk'si:gtə], Partizip II: [buk'si:gt]

Hörbeispiele: —

Bedeutungen:

[1] *transitiv, Seefahrt*: ein Schiff durch **Lotsen** ins **Schlepptau** nehmen und an einen bestimmten Ort schleppen

[2] *transitiv, umgangssprachlich*: mit Mühe an einen anderen Ort bringen

[a] Dinge

[b] Menschen, auch Tiere

Herkunft:

Lehnwort des 17. Jahrhunderts zunächst in den Schreibweisen *buxiren*, *buchsieren* aus dem **Niederländischen** *boegseren*, weiterentwickelt aus den alten Formen *boesjaarden*, *boesjaren*, kombiniert mit dem niederländischen Wort *boeg* (**Bug**); die niederländischen Urformen gehen auf das **portugiesische** *puxar* (ziehen, schleppen) zurück, das sich seinerseits aus dem lateinischen *pulsare* (stoßen) ableitet; die jetzige Schreibweise stammt aus dem 19. Jahrhundert ^[1]

Synonyme:

[1] geleiten, lotsen, ins Schlepptau nehmen, schleppen

[1,2a] drücken, schieben, ziehen

[2a] schleifen, stoßen

[2b] dirigieren, drängen, unter die Arme greifen, schubsen, weiterhelfen

Gegenwörter:

[2a] karren, weggrollen, werfen

Berechnung von Semantik

Folksonomy / Social Tagging

Eine durch Benutzer erzeugte Stichwortsammlung zu einer Internetseite, Bild oder Artikel.

- + Sehr genau, da menschlich generiert
- Kein Algorithmus und somit nicht generisch anwendbar
- Kein kontrolliertes Vokabular
- Manipulierbar



Quellen:

Delicious.com, mister-wong.de (jeweils Screenshots der TagCloud)

Berechnung von Semantik

Clustern

Indem man z. B. die Wörter in den Titeln der ersten 100 Treffern einfach mathematisch gruppiert.

- + rel. schnell
- Ungenau (da nicht semantisch, sondern eben nur geclustert)

The screenshot shows the Clusty search interface. At the top, there's a navigation bar with links for 'web', 'news', 'images', 'wikipedia', 'blogs', 'jobs', and 'more'. A search bar contains the query 'seo'. Below the search bar, there are tabs for 'clusters', 'sources', and 'sites'. The 'clusters' tab is active, showing a list of clusters. The 'Design' cluster is selected, containing 24 documents. The main content area displays a list of search results for the 'Design' cluster, including titles like 'SEO Design Solutions™ Proven SEO Results!', 'Clever SEO', 'Rankings of Best SEO, PPC, Web Design and Development', 'Search Engine Optimization by SEO Design Solutions™', and 'SEO Expert'.

Bild:
www.clusty.com (Vivisimo)

Berechnung von Semantik

HAL (Hyperspace Analogue to Language)

Jedes Wort wird durch die Gesamtheit seiner Nachbarschaften im Kontext repräsentiert, oder einfacher ausgedrückt: **Wörter mit ähnlicher Bedeutung erscheinen in ähnlichen Sätzen.**

- + Semantische Nähe
- = Ergebnisse ähnlich LSI

Wir analysieren ein paar Texte und stellen ein **Kookkurrenzen** fest:
(sprich „a“ findet sich oft im Zusammenhang mit „b“ genannt)

	mountain	valley	river	mouse	cat	dog
mountain	0	12	8	0	0	0
valley	12	0	9	0	0	0
river	8	9	0	0	0	0
mouse	0	0	0	0	9	0
cat	0	0	0	9	0	5
dog	0	0	0	0	5	0

Naive HAL Matrix

HAL ← Kevin Lund and Curt Burgress ,1996 <http://www.psychonomic.org/search/view.cgi?id=1105>

Berechnung von Semantik

LSI – Latent Semantic Indexing

1) Man bilde die Term-Dokumenten-Matrix aller URLs

Sample Term by Document matrix

	<i>access</i>	<i>document</i>	<i>retrieval</i>	<i>information</i>	<i>theory</i>	<i>database</i>	<i>indexing</i>	<i>computer</i>
Doc 1	x	x	x			x	x	
Doc 2				x*	x			x*
Doc 3			x	x*				x*

2) lege darüber den semantischen Raum

(via Singularitätswertzerlegung der wichtigsten Konzepte/Wörter)

(Idealerweise mit vorhanden Synonymdatenbanken noch weiter verkleinern)

➔ Die Matrix wird kleiner , lässt sich schneller rechnen,

Berechnung von Semantik

LSI – Latent Semantic Indexing

	Dokumente Lafontaine	Schröder	Euro	Uefa	Beckenbauer
d1	1	1	1	0	0
d2	2	2	2	0	0
d3	1	1	1	0	0
d4	5	5	5	0	0
d5	0	0	0	2	2
d6	0	0	0	3	3
d7	0	0	0	1	1

Die SVD ermittelt daraus:

$$U + \begin{matrix} 0.18 & 0.00 \\ 0.36 & 0.00 \\ 0.18 & 0.00 \\ 0.90 & 0.00 \\ 0.00 & 0.53 \\ 0.00 & 0.80 \\ 0.00 & 0.27 \end{matrix} \quad D + \begin{matrix} 9.64 & 0.00 \\ 0.00 & 5.29 \end{matrix} \quad V^T + \begin{matrix} 0.58 & 0.58 & 0.58 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.71 & 0.71 \end{matrix}$$

Intuitiv enthält die Dokumentenkollektion also zwei Themen, nämlich Politik und Sport.

Ein weiteres Dokument $d8 = (1 \ 1 \ 0 \ 0 \ 0)$ mit dem einzigen Feature "Lafontaine" würde dann auf den **Themenvektor** $d8 \times V = (0.54 \ 0)$.

Eine Query $q = (0 \ 0 \ 1 \ 0 \ 0)^T$, die nach dem Feature "Euro" sucht, würde auf den Themenvektor $V^T \times q = (0.58 \ 0)^T$ abgebildet. Unabhängig vom wirklich verwendeten Ähnlichkeitsmaß im Themenvektorraum wäre also d8 vermutlich ein sehr guter Treffer für q.

Quelle:
www.mpi-inf.mpg.de

Berechnung von Semantik

PLSI (Probabilistic Latent Semantic Indexing) ← Thomas Hofmann 1999

Im Vergleich zu LSI welches auf einer Matrixzerlegung basiert, hat die probabilistische Variante statistische Grundlagen (**bedingte Wahrscheinlichkeit**), um eine höhere Präzision zu erreichen. Die Dimensionsreduktion erfolgt nicht via SVD (Singularitäts-Wert-Zerlegung), sondern auf **Bayes** beruhende Wahrscheinlichkeitsberechnungen.

HTMM (Hidden Topic Markov Model) ← Amit Gruber 2007

- Annahme: alle Wörter in einem Satz haben das gleiche Thema.
 - Annahme: nachfolgende Sätze haben das gleiche Thema.
 - „Latent“, sprich das eigentliche Thema ist aber versteckt.
 - Anwendung des „Hidden Markov Modells“ zur **Mustererkennung verborgener Zustände**.
- = Hidden Topic Markov Modell

(verborgene Zustände von DNA-Sequenzen, 23andme ??)

Berechnung von Semantik

Der „semantischen Raum“

- 1) **Syntagmatischen** Assoziationen von Wörtern
→ Kookkurrenzen in Texten eines Korpus

- 2) **Paradigmatischen** Assoziationen
→ Vergleichs der Kontexte der Vorkommen der Wörter

- 3) **Bedeutungspunkte**
→ zu jedem Wort einen Punkt in einem hochdimensionalen Vektorraum

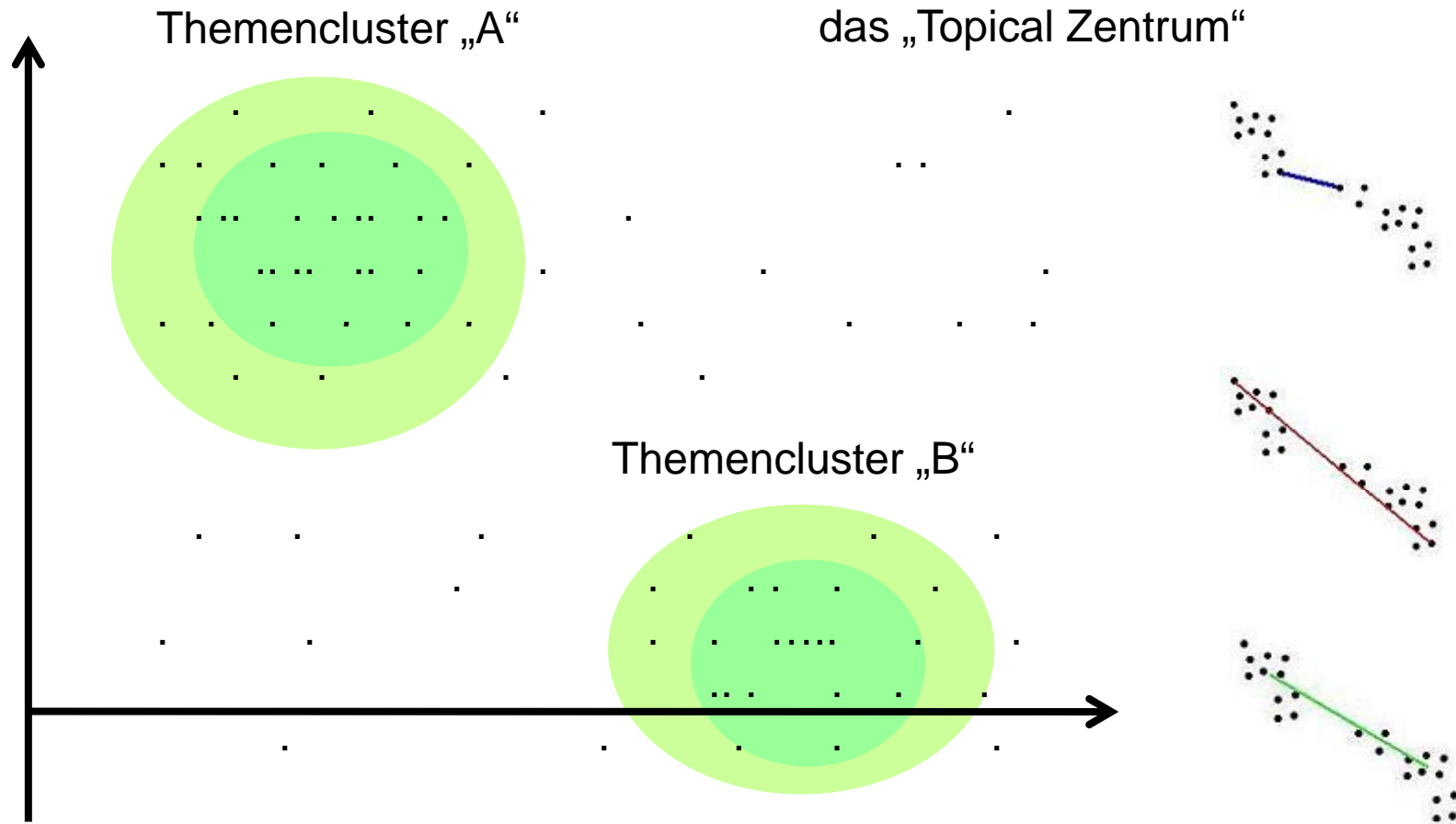
- 4) **Semantischer Raum**
→ Menge der Bedeutungspunkte, topologische Nachbarschaften
→ räumliche Lage zueinander bestimmt Bedeutungsähnlichkeiten

Die Bedeutung eines Wortes nicht isoliert beschreibbar, sondern wird wie durch die Koordinaten eines Punktes durch die Beziehungen zu allen anderen Wörtern bestimmt.

Berechnung von Semantik

Bedeutungspunkte → semantischer Vektorräume

z.B. „Gravitationszentrum“:
innerhalb einer Topical Community
das „Topical Zentrum“



Berechnung von Semantik

Beispiel: automatische Sortierung von Google News in eine Kategorie (LSI-Vektoren)

The screenshot shows the Google News Germany interface. At the top, there is a search bar with the text "Google news Deutschland" and buttons for "News-Suche" and "Das Web durchsuchen". Below the search bar, there is a navigation menu with categories: Schlagzeilen, International, Deutschland, Wirtschaft, Wissen/Technik, Unterhaltung, Sport, Gesundheit, Panorama, and Meistgeklickt. The "Schlagzeilen" category is selected. The main content area displays two news items. The first item is titled "Viktoria Rebensburg mit Vollgas zum Olympiasieg" and is dated "WELT ONLINE - Vor 37 Minuten". The second item is titled "Käßmann-Rücktritt: 'Jeder Karnevalist hätte seine Scherze gemacht'" and is dated "STERN.DE - Vor 1 Stunde". Both items have a link to "Alle 459 Artikel »" and "Per E-Mail senden". The links "Alle 459 Artikel »" and "Alle 3.117 Artikel »" are circled in red.

Deutschland News-Suche Das Web durchsuchen

Deutschland

Schlagzeilen

Viktoria Rebensburg mit Vollgas zum Olympiasieg
WELT ONLINE - Vor 37 Minuten
Von Jens Hungermann 25. Februar 2010, 22:00 Uhr Nach ihrem überraschenden Olympiasieg im Riesenslalom kannte der Jubel bei Viktoria Rebensburg keine Grenzen mehr. Mit 20 Jahren ist sie die zweitjüngste alpine Goldmedaillengewinnerin, die Deutschland je ...
[Olympiasiegerin Rebensburg: «Fantastisch»](#) sueddeutsche.de
[Rebensburg: Eltern verpassen Riesenslalom-Gold-Lauf](#) BILD
[Derwesten.de](#) - [Hamburger Abendblatt](#) - [STERN.DE](#) - [Tagesspiegel](#)
Alle 459 Artikel »

Käßmann-Rücktritt: "Jeder Karnevalist hätte seine Scherze gemacht"
STERN.DE - Vor 1 Stunde
Im Interview mit stern.de spricht Günther Beckstein über den Rücktritt von Margot Käßmann, die Versuchung Alkohol und die Zukunft der evangelischen Kirche. Günther Beckstein Der ehemalige bayerische Ministerpräsident Günther Beckstein (CSU) ist ...
[Beckstein zu Käßmann: „Einem Mann hätte man eher verzeihen“](#) FOCUS Online
[Käßmanns-Rücktritt: Ratlos vor der großen Lücke](#) Tagesspiegel
[FAZ](#) - [Frankfurter Allgemeine Zeitung](#) - [WELT ONLINE](#) - [Derwesten.de](#) - [Augsburger Allgemein](#)
Alle 3.117 Artikel »

Deutschland

›

Auf der Suche nach dem semantischen Raum

LSO – Latent Semantische Optimierung

Suche nach dem „semantischen Raum“

Wort: evangelisch

Anzahl: 109

Häufigkeitsklasse: 15 (d.h. *der* ist ca. 2^{15} mal häufiger als das gesuchte Wort)

Morphologie: evangelisch

Grammatikangaben: Wortart: Adjektiv

Pragmatikangaben: etym: griech.

Relationen zu anderen Wörtern:

- Synonyme: [protestantisch](#)
- wird referenziert von: [protestantisch](#)

Links zu anderen Wörtern:

- Grundform: [evangelisch](#)
- Haupteintrag (Polysem): [evangelisch](#), [evangelisch](#)
- ist ein(e) [protestantisch](#)
- Form(en): [evangelischen](#), [evangelische](#), [evangelischer](#), [evangelisch](#), [evangelischem](#), [evangelisches](#)
- Untereinträge: [evangelisch](#), [evangelisch](#)

Dornseiff-Bedeutungsgruppen:

- 22.1 Religiosität, Glaube: [evangelisch](#), [gläubig](#), [gottergeben](#), [gottesfürchtig](#), [gottgefällig](#), [heilig](#), [kanonisch](#), [katholisch](#), [orthodox](#), [protestantisch](#), [puritanisch](#), [rein](#), [religiös](#), [tugendhaft](#), [unschuldig](#)
- 22.12 Gebet, Frömmigkeit: [biblisch](#), [eingegeben](#), [erbaulich](#), [evangelisch](#), [geheiligt](#), [göttlich](#), [inspiriert](#), [prophetisch](#)

Beispiel(e):

Nur eins macht mir Sorge: Angela Merkel ist **evangelisch**. (Quelle: [fr-aktuell.de vom 08.01.2005](#))

Die 54-jährige Juristin ist in Frankfurt geboren, wohnt seit 21 Jahren in Neu-Isenburg, ist **evangelisch**, verheiratet und hat zwei erwachsene Söhne. (Quelle: [fr-aktuell.de vom 13.01.2005](#))

Moshammer war **evangelisch**, trat jedoch aus der Kirche aus. (Quelle: [n-tv.de vom 19.01.2005](#))

[weitere Beispiele](#)

Signifikante Kookkurrenzen für evangelisch:

[katholisch](#) (506), [getauft](#) (55), [Kirche](#) (55), [Krippen](#) (53), [Krippentraditionen](#) (45), [Taufunterlagen](#) (45), [jüdisch](#) (41), [getaufter](#) (38), [Heimatemuseum](#) (35), [geboren](#) (30), [geschieden](#) (30), [konfessionell](#) (29), [Oelsner](#) (27), [muslimisch](#) (27), [herumgeprägten](#) (26), [Katholiken](#) (25), [Einwohner](#) (25), [draufsteht](#) (24), [Kindheit](#) (23), [Evangelium](#) (23), [Seid](#) (23), [Religionszugehör](#)

Quelle:

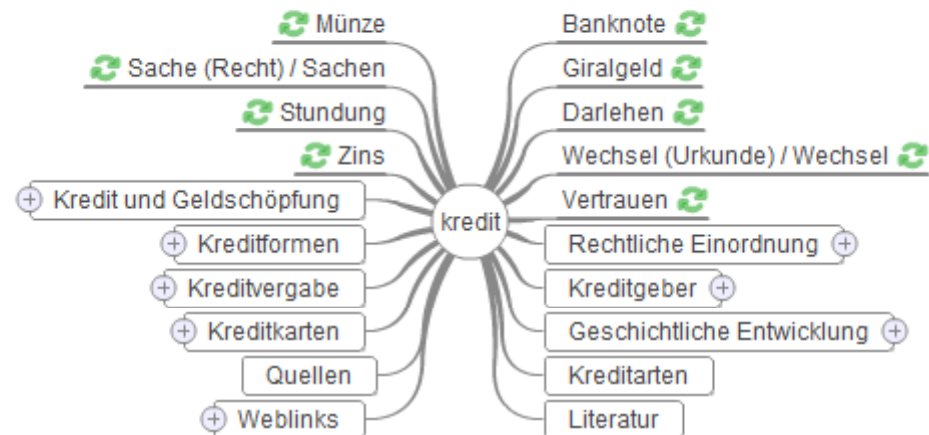
Wortschatz Uni-Leipzig

LSO – Latent Semantische Optimierung

Suche nach dem „semantischen Raum“

wiki mindmap

Select a Wiki: Enter your Topic:



LSO – Latent Semantische Optimierung

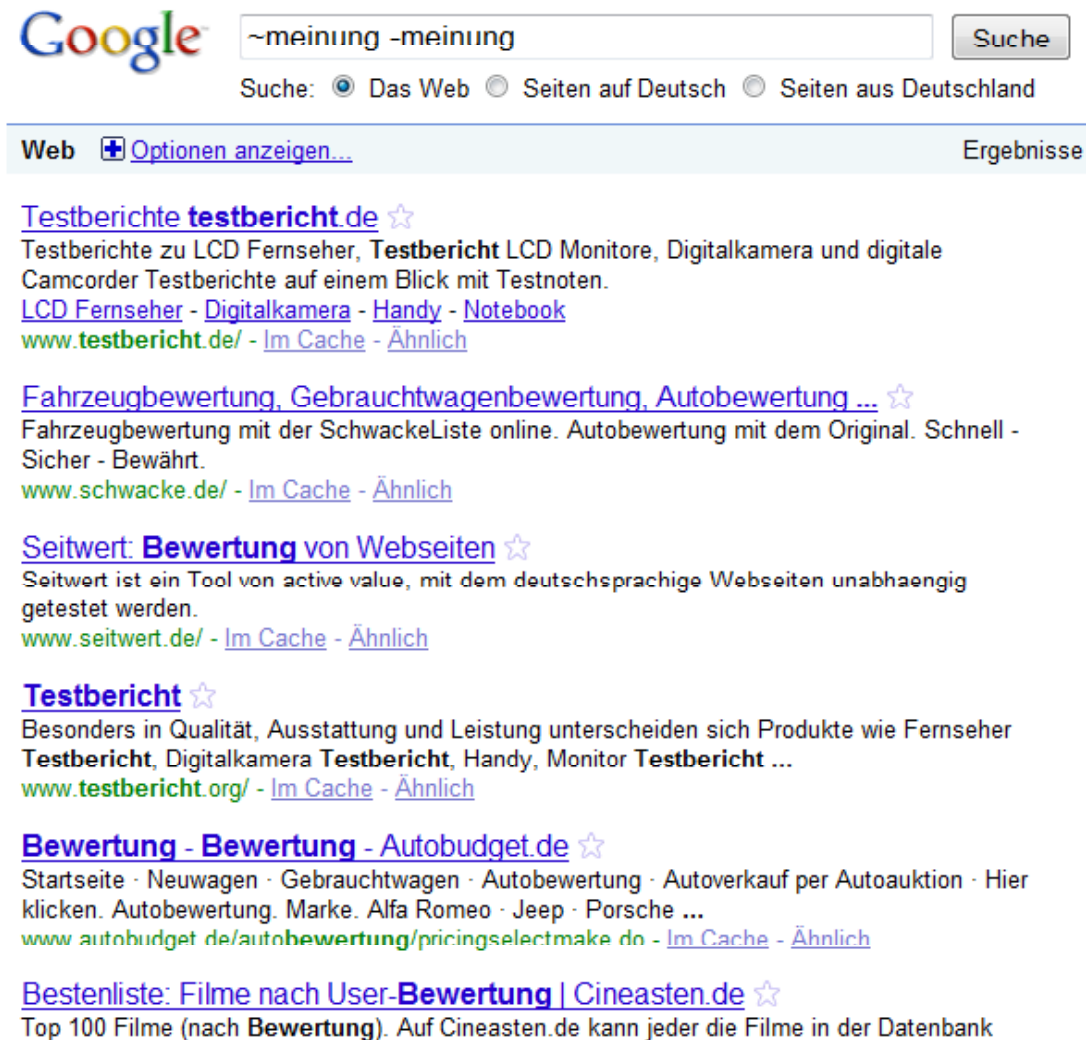
Suche nach dem „semantischen Raum“



Predicted Items
evangelische
kirche
evangelisch
protestantismus
katholische
gesellschaft
homepages
reformation
diakoniestation
ehelosigkeit
zölibat

LSO – Latent Semantische Optimierung

Suche nach dem „semantischen Raum“



Google

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web Ergebnisse

[Testberichte testbericht.de](#) ☆
Testberichte zu LCD Fernseher, **Testbericht** LCD Monitore, Digitalkamera und digitale Camcorder Testberichte auf einem Blick mit Testnoten.
[LCD Fernseher](#) - [Digitalkamera](#) - [Handy](#) - [Notebook](#)
www.testbericht.de/ - [Im Cache](#) - [Ähnlich](#)

[Fahrzeugbewertung, Gebrauchtwagenbewertung, Autobewertung ...](#) ☆
Fahrzeugbewertung mit der SchwackeListe online. Autobewertung mit dem Original. Schnell - Sicher - Bewährt.
www.schwacke.de/ - [Im Cache](#) - [Ähnlich](#)

[Seitwert: Bewertung von Webseiten](#) ☆
Seitwert ist ein Tool von active value, mit dem deutschsprachige Webseiten unabhängig getestet werden.
www.seitwert.de/ - [Im Cache](#) - [Ähnlich](#)

[Testbericht](#) ☆
Besonders in Qualität, Ausstattung und Leistung unterscheiden sich Produkte wie Fernseher **Testbericht**, Digitalkamera **Testbericht**, Handy, Monitor **Testbericht** ...
www.testbericht.org/ - [Im Cache](#) - [Ähnlich](#)

[Bewertung - Bewertung - Autobudget.de](#) ☆
Startseite · Neuwagen · Gebrauchtwagen · Autobewertung · Autoverkauf per Autoauktion · Hier klicken. Autobewertung. Marke. Alfa Romeo · Jeep · Porsche ...
www.autobudget.de/autobewertung/pricingselectmake.do - [Im Cache](#) - [Ähnlich](#)

[Bestenliste: Filme nach User-Bewertung | Cineasten.de](#) ☆
Top 100 Filme (nach **Bewertung**). Auf Cineasten.de kann jeder die Filme in der Datenbank

LSO – Latent Semantische Optimierung

Semager Keyword API

All In- and Output UTF-8 encoded

Query

Language (optional, default: de)

de ▾

Output (optional, default: xml)

xml ▾

Count (optional, max. number of results, default: 20)

Threshold (optional, min. correlation to query in percentag

Sort (optional, sort after the correlation to query or absolut
(Freq. means its sorts after how much common a word is)

correlation ▾

Daten absenden

```
-<results>
- <resultset title="Semager: Keyword API" query="seo" spell="" lang="de"
- <item>
  <score>83</score>
  <tag>suchmaschinen optimierung</tag>
</item>
- <item>
  <score>76</score>
  <tag>sem</tag>
</item>
- <item>
  <score>76</score>
  <tag>webseiten</tag>
</item>
- <item>
  <score>70</score>
  <tag>suchmaschinen</tag>
</item>
- <item>
  <score>70</score>
  <tag>optimierung</tag>
</item>
- <item>
  <score>70</score>
  <tag>suchmaschinenoptimierung</tag>
</item>
```

LSO – Latent Semantische Optimierung

Semager URL Analyse

Geben Sie hier Ihre Internetseite ein:

Suchbegriff(e) auf die Sie die Webseite optimiert haben (falls vorhanden, für LSO-Tool):

Webseiten Analyse

Tools

- [Webseite analysieren](#)
- [Webmaster-Tools](#)
- [Textduplikate finden](#)
- [Semantic Business](#)

About

- [Semager Blog](#)
- [Feedback](#)

Die wichtigsten Wörter aus Sicht einer Suchmaschine

Übereinstimmung	Keyword	Verwandte Keywords dazu
95%	seo	Nachschlagen
95%	berlin	Nachschlagen
94%	campixx	Nachschlagen
81%	unkonferenz	Nachschlagen
67%	suchmaschinenoptimierung	Nachschlagen
61%	genannte	Nachschlagen
59%	schwerpunkt	Nachschlagen
58%	informationsaustausch	Nachschlagen
49%	spaß	Nachschlagen
47%	location	Nachschlagen
42%	ziel	Nachschlagen
39%	ideen	Nachschlagen
39%	leute	Nachschlagen
38%	festen	Nachschlagen
38%	referenten	Nachschlagen
36%	beteiligen	Nachschlagen
34%	hotel	Nachschlagen
34%	müggelsee	Nachschlagen
33%	barcamps	Nachschlagen

Latent Semantische Analyse

Kategorie

19% Computer und Internet - Webdesign / Webmaster
19% Computer und Internet - Infosuche
14% Computer und Internet - Programmierung

LSO Optimierungsgrad

Davon ausgehend, das auf „seo“ „berlin“ „campixx“ optimiert worden ist.

29.09% Übereinstimmung

Keywords die aus semantischer Sicht ergänzt werden könnten

94% berliner	
85% berlins	nicht vorhanden
84% suchmaschinen optimierung	nicht vorhanden
82% deutschland	nicht vorhanden
79% brandenburger	nicht vorhanden
78% friedrichshain	nicht vorhanden

Meta

Statistik:

Größe der Seite: 12.5 KB
Ohne Scripte, CSS: 12.3 KB
Ohne HTML: 2.5 KB

Anzahl aller sichtbaren Wörter: 248
Abzüglich Stopwords: 92
Davon eindeutig: 82

Zeichensatz: ISO-8859-2

HTTP-Header: -
Meta-Tag: WINDOWS-1252
Content: ISO-8859-2

Sprache: de

HTTP-Header: -
Meta-Tag: de
Content: de
Encoding: (many)

Ladezeit:

DNS-Lookup:	0.00004 sec
Connect to Host:	0.009 sec
Download-Transfer:	0.074 sec
Gesamt Transfer:	0.127 sec

LSO – Latent Semantische Optimierung

Antworten

LSO ist

- mit verwandten Keywords ranken
- nachhaltig, da zukunftsweisend
- Keywordrecherche
- Onpage Optimierung

Was muss getan werden?

- Nicht immer das gleiche Keyword verwenden, sondern auch Wörter aus dem Umfeld.

Synonyme bei der Google Suche abschalten

- Indem man ein „+“ vor das Wort hängt

Synonyme bei der Google Suche finden

- ~Keyword –Keyword (was ist bold?)

Glaskugel

[Erweiterte Suche](#)

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web [+ Optionen anzeigen...](#)

Ergebnisse 1 - 10 von ungefähr 46.600.000

Meinten Sie: [internettelefonie](#) Die ersten 2 angezeigten Ergebnisse

[Skype – Kostenlose Anrufe und preiswerte Internettelefonie](#)

Anrufe zwischen Skype-Nutzern sind immer kostenlos, aber mit dem im Voraus bezahlten Skype-Guthaben und den monatlichen Abonnements können Sie Freunde und ...

www.skype.com/intl/de/ - [Im Cache](#) - [Ähnlich](#)

[Internettelefonie - Telefonie über das Internet](#)

Die **Internettelefonie** (VoIP) ermöglicht die Telefonie über das Internet mit einem Telefon und kann den Festnetz-Telefonanschluss nahezu ersetzen.

www.telespiegel.de/.../internettelefonie_-_alle_detai.html - [Im Cache](#) - [Ähnlich](#)

Fake (noch) ;-)

DANKE !

Wie finde ich den ContentDNA zu einem Wort / Thema heraus?

www.semager.de/keywords/

das ist ein Affiliate Link ;-)

bis zur SEO-Campixx oder heute Abend auf der Pubcon ☺

Appendix

APPENDIX

Themenrelevante Links

Zitat „boeser SEO“:

Mark: Mich würde ein Vergleich von themenrelevanten contra themenfremden Links interessieren.

„ Totaler Bullshit... Aber es gehört ja zum guten SEO Ton [...] [...] wahrscheinlich das Google da in naher Zukunft den Algo besser anpassen wird [...] “

APPENDIX

Themenrelevante Links

Zitat „Sistrix“:

*[...] dass **die derzeitigen Google-SERPs so nicht funktionieren***

*[...] im **derzeitigen Ranking kann ich noch keine signifikanten Auswirkungen davon erkennen***

[...] das sie mit Neuerungen lieber warten, bis Yahoo oder Microsoft den Abstand etwas verkürzt hat

*[...] Themenrelevanz bei den Verlinkungen wird in Zukunft wichtig sein und **wer heute für zukünftige Projekte nicht auf themennahe Links achtet, investiert seine Zeit ungeschickt***

APPENDIX

Synonyme natürlich verwenden

Zitat “Matt Cutts”:

Think about the different words that searchers might use when looking for your content. Don't just use technical terms – think about real-world terms and slang that users will type. For example, if you're talking about a “usb drive,” some people might call it a flash drive or a thumb drive. Bear in mind **the terms that people will type and think about synonyms that can fit naturally into your content.** Don't stuff an article with keywords or make it awkward, but if you can incorporate different ways of talking about a subject in a natural way, that can help users.

Quelle:

<http://www.mattcutts.com/blog/google-synonyms/>

APPENDIX

Wer hat`s erfunden?

- Termvektor: The terms of an HTML page are sequences of non-space characters found by filtering and normalizing the page's *term candidates*.
Raymie Stata, Krishna Bharat und Farzin Maghoul:
The Term Vector Database: fast access to indexing terms for WebPages
- Themenvektor: Termvektor vieler inhaltlich gleicher Webseiten
Vordefiniert, evtl. auch via DMOZ oder Yahoo-Verzeichnis
Krishna Bharat und Monika Henzinger:
Improved Algorithms for Topic Distillation in a Hyperlinked Environment
- Thema: Themenvektor mit dem höchsten Grad an Übereinstimmung aus
Vergleich zwischen Themenvektor dieser Webseite und
vordefinierten Themenvektoren

APPENDIX

Literatur

Doktorarbeit: Nachbarschaften im semantischen Raum

<http://ubt.opus.hbz-nrw.de/volltexte/2006/373/pdf/Wegner-NachbarschaftenImSemantischenRaum.pdf>

Magisterarbeit: Textverstehen als Prozess assoziativen Schließens

http://www.kognitioninbielefeld.de/fileadmin/user_upload/Documents/dokumentmag.pdf

Englischsprachige Untersuchung

<http://www.latentsemanticoptimization.com/>

LSO Experiment

<http://www.h1h2h3.de/auswertung-zum-keywordtest-sportaktivitaeten/>