Frameworks zur Entwicklung von Suchmaschinen

Dipl.-Inf. Frank Hofmann

Potsdam

1. Juli 2007

Zielsetzung

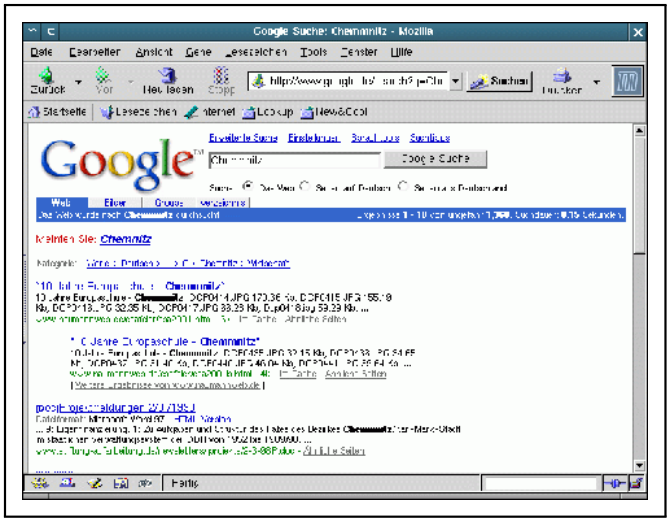
Fragen, die sich jeder stellt

- Wie funktioniert eine Suchmaschine?
- Warum finde ich eigentlich nicht das, was ich suche?
- Suchen geht mit Google. Und womit noch?
- Kann ich mir auch meine eigene Suchmaschine bauen?

Übersicht

1. Betrachtungen zum Suchprozess
2. Erweiterte Suche
3. Frameworks/Projekte/Libraries
4. Projekte zur Visualisierung
5. Referenzen

Suchmaschinen - eine Begriffsklärung

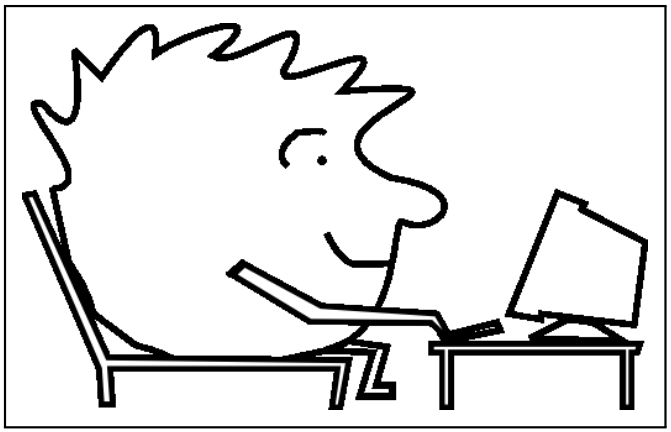


Ziel:
Informationen anzeigen, nach denen ich suche

Zugehörigkeit zu

- Informationssystemen
- Information Retrieval Systemen

Der Informationssuchende (1)



- Person, die eine Information oder Auskunft benötigt
- Person stellt Anfrage an Informationssystem, um Auskunft zu erhalten
- erwartet auf seine Anfrage eine möglichst exakte, vollständige und für ihn verständliche Antwort in möglichst kurzer Zeit

Der Informationssuchende (2)

Probleme und Fragen

- Auswahl der geeigneten Informationsquelle (welche Quelle deckt meinen Informationsbedarf am besten?)
- Formulieren einer Anfrage (wie versteht das System meine Frage?)
- Optimale Ergebnisse setzt detaillierte Kenntnisse des Systems und die Ablageform der Daten voraus (was wird überhaupt gespeichert?)
- Begutachtung des Suchergebnisses (werden meine Erwartungen erfüllt? Falls nicht, wie formuliere ich um?)
- Sucht das IS nur im Rechercheergebnis, oder wieder über die gesamte Datenbasis? (Verfeinerung vs. Überblick)

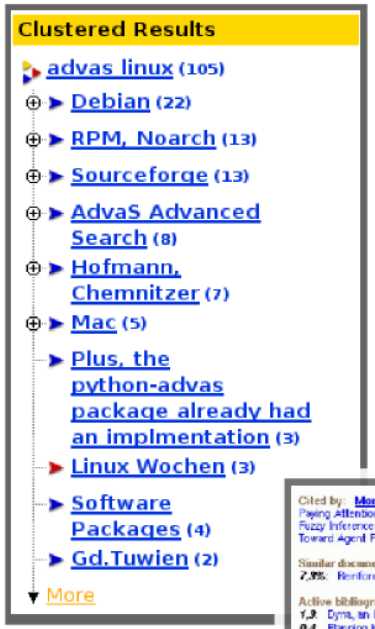
Der Informationssuchende – Subjektive Auswahlkriterien

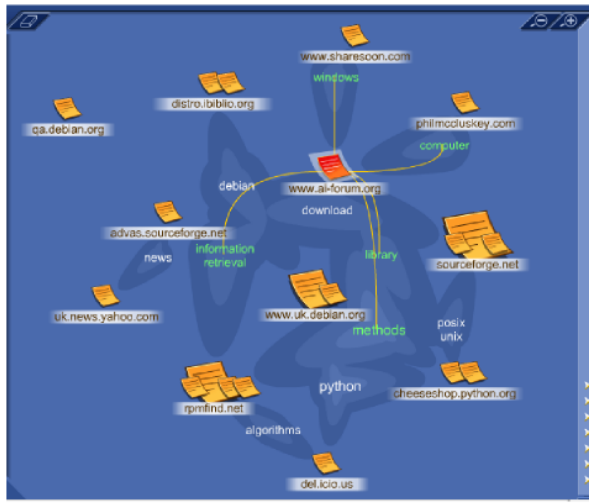
Entscheidung über potentielle Relevanz eines Ergebnisses:

- (Daten)Format
- Gestaltung, Layout
- Bücher: Geruch, Papiersorte, Einband
- kennt oder schätzt den Autor
- Anspruch, Informationslevel

Bestandteile des Suchprozesses

Formulierung der Anfrage > Verarbeitung der Anfrage > Ausgabe des Ergebnisses





Bewertung des Ergebnisses

Kriterien:

- Anzahl der Dokumente
- Qualität der Informationen
  - Relevanz
  - Präzision
  - Fallout
- Antwortzeit
- Betriebsaufwand
- Nutzerfreundlichkeit

Erweiterte Suche – Begriff, Sinn und Zweck

- alle Verfahren, die über das einfache Vergleichen und Aufsuchen von vollständigen Worten in einem Text hinausgehen (exakte Muster)
- mit Vorverarbeitung (Indexierung, Katalogisierung und Zuordnung zu einer bestimmten Thematik)
- Verwendung von Deskriptoren (Begriffen, die das Dokument inhaltlich beschreiben)

Verfahren zur Erzeugung von Deskriptoren (1)

Je häufiger ein Wort in einem Text vorkommt, umso relevanter ist der Text in Bezug auf dieses Wort.

Das ist unser Text, den wir als einfaches Beispiel nehmen. Unser Text zeigt, welche statistischen Verfahren benutzt werden.

Statistik		
Wort	Anzahl	Stopliste
das	1	ja
ist	1	ja
unser	2	nein
Text	2	nein
den	1	ja
...		

Verfahren zur Erzeugung von Deskriptoren (2)

Verfahren aus der Sprachwissenschaft

- Morphologische Analyse der Wortformen, Stemming (Zurückführung auf grammatikalische Grundform)
- Semantische Analyse des Wortes (Betrachtung der Bedeutung, Suche nach Synonymen oder Antonymen)

Wortform	Wortstamm
fliegen, flog, geflogen, Flieger	flieg
schwimmen, schwamm, geschwommen, Schwimmer	schwimm

Verfahren zur Erzeugung von Deskriptoren (3)

Algorithmen für ähnlich klingende Worte

- Umwandlung in einen Buchstabencode
- SOUNDEX, Metaphone, NYIIS, Caverphone

	soundex	metaphone	nyiis	caverphone
schmidt :	S253	sxmtt	sssnad	SKMT111111
schmid :	S253	sxmt	sssnad	SKMT111111
schmitt :	S253	sxmt	sssnatt	SKMT111111
smith :	S530	sm0h	snatt	SMT1111111
smythe :	S530	smy0h	snatt	SMT1111111
schmied :	S253	sxmt	sssnaad	SKMT111111
mayer :	M600	myr	naaar	MA11111111
meier :	M600	mr	naaar	MA11111111
maier :	M600	mr	naaar	MA11111111
meyer :	M600	myr	naaar	MA11111111

Verfahren zur Erzeugung von Deskriptoren (4)

Textkorrektur

- Rechtschreibprüfung
  `aspell`, `pspell` („portable spell checker interface library“)
- Nutzung der WordNet-Bibliothek
  ermöglicht Untersuchung eines Wortes hinsichtlich seines Vorkommens in der Sprache
- Fremdwörterbuch
- etymologische Wörterbücher
  Berücksichtigung historischer Schreibweisen, Lautveränderung
- Spracherkennung (über Wortverteilung)

Deskriptor-basierte Verfahren

Varianten

Variante	Beispiel
Term Weighting Gewichtung von Deskriptoren	`Python (0.5) image (0.3) Linux (0.2)`
Boolesche Operatoren Klammerung, Verknüpfung der Suchbegriffe mit AND, OR etc.	`(Python AND image) AND (Linux OR Mac)`
Reguläre Ausdrücke	`/[Pp]ython.[Ii]maging library/`

APACHE-Projekt: Lucene

`http://lucene.apache.org`

- Beschreibung:
  „Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform“
- Einsatzbereich:
  Web- und Intranetserver

APACHE-Projekt: Nutch

`http://www.apache.org/nutch`

- baut auf Lucene auf
- erweitert durch Web-Crawler, Datenbank für Link-Graphen
- Parser für HTML, Plain Text, JavaScript, Microsoft Power Point, Microsoft Word, PDF, RSS, RTF, MP3 und ZIP
- realisiert durch Plugins

APACHE-Projekt: Hadoop

`http://www.apache.org/hadoop`

- baut auf Lucene auf
- für verteiltes Rechnen (Cluster)
- Prozesse zerlegbar auf einzelne Knoten
- verteiltes Dateisystem: Hadoop Distributed Filesystem (HDFS)

APACHE-Projekt: Solr

`http://www.apache.org/solr`

- Erweiterung von Lucene
- „high performance search server“
- verfügt über
  - Schnittstellen zu XML/HTTP und JSON/Python/Ruby
  - Hervorhebung der Suchterms im Ergebnis
  - Caching und Replikation

advas Advanced Search

`http://advas.sourceforge.net`

- basiert auf Python
- bisher released als RPM, seit Ende 2006 Teil von Debian Etch
- Ziel
  - Schaffung einer Bibliothek für Suchverfahren
  - Nutzung für Lehr-, Lern- und Testzwecke
- Anwendungsbereich: lokale Suche

Ferret

`http://ferret.davebalmain.com/trac/`

- basiert auf Ruby
- eine high-performance, full-featured Library zur Textsuche
- Voraussetzungen
  - Ruby 1.8
  - C-Compiler und Make
- Anwendungsbereich: lokale Suche

KinoSearch

http://www.rectangular.com/kinosearch/

- basiert auf Perl
- Suchmaschinenbibliothek
- Anwendungsbereich: lokale Suche, Webserver

iglu-java

`http://iglu-java.sourceforge.net`

- basiert auf Java
- Suchmaschinenbibliothek
- Anwendungsbereich: lokale Suche, Webserver

beagle

`http://www.beagle-project.org`

- Suchmaschinenbibliothek
- Anwendungsbereich: lokale Suche
- Real-Time-Applikation (im laufenden Betrieb)
- kann nahezu alle lokalen Daten indexieren (Mails, Bilder, Dokumente, Webseiten)
- integriert in KDE und GNOME

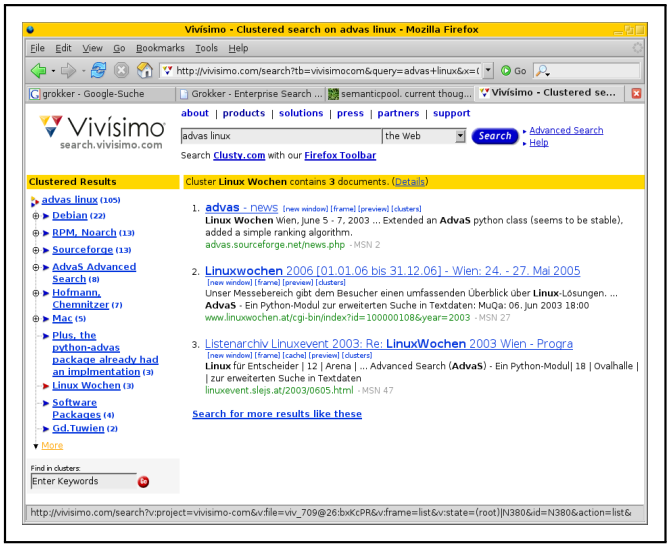
ht://Dig

`http://www.htdig.org/`

- Suchmaschinenprogramm
- Anwendungsbereich: WWW, Intranet
- indexiert Webdaten

vivisimo

`http://vivisimo.com`



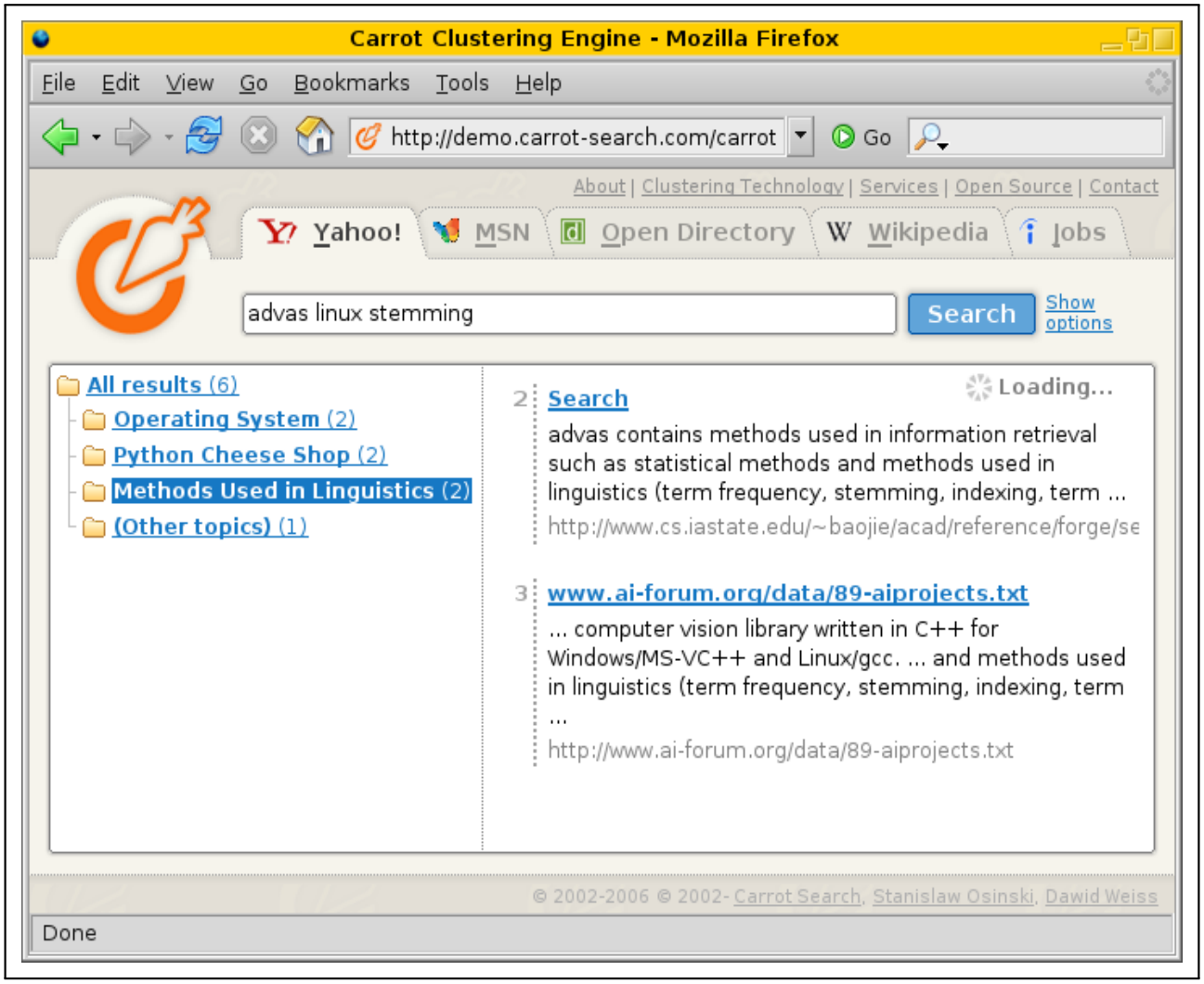
- Nutzung verschiedener Suchdienste
- thematische Gruppierung der Ergebnisse

carrot, carrot2 (1)

`http://project.carrot2.org/` (Open Source Library)
`http://company.carrot-search.com` (kommerzielle Variante)

- Nutzung verschiedener Suchdienste
- Thematische Gruppierung der Ergebnisse

carrot, carrot2 (2)

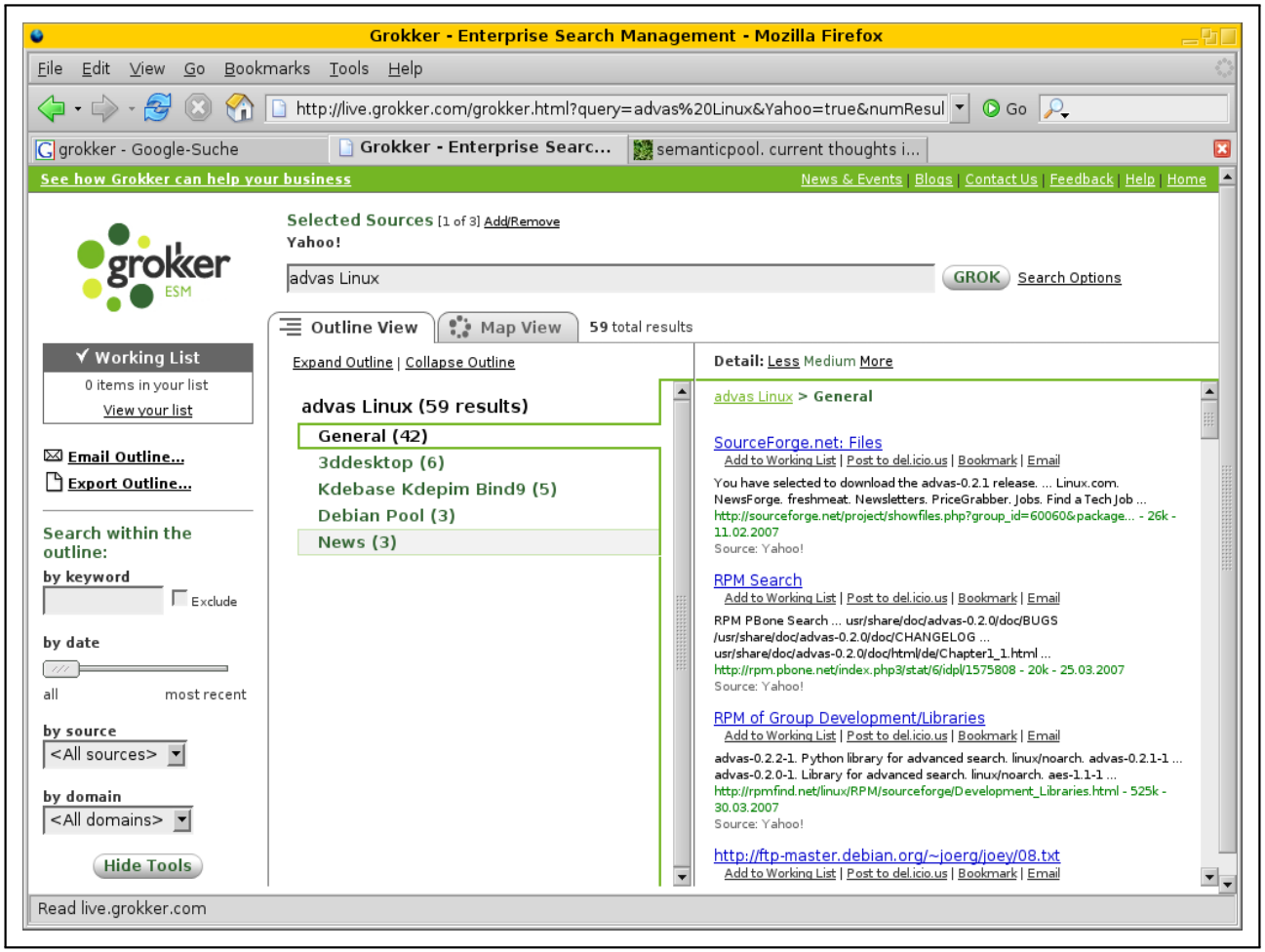


grokker (1)

`http://www.grokker.com`

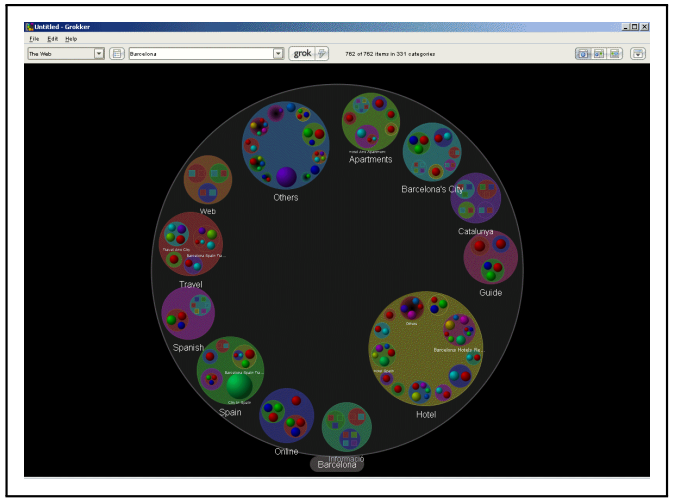
- Visualisierung von Suchergebnissen
- Zugriff auf verschiedene Suchmaschinen (z.B. Google)
- zwei Darstellungen der Suchergebnisse: thematische Gruppierung (HTML) grafisch in lustigen „Blasen“

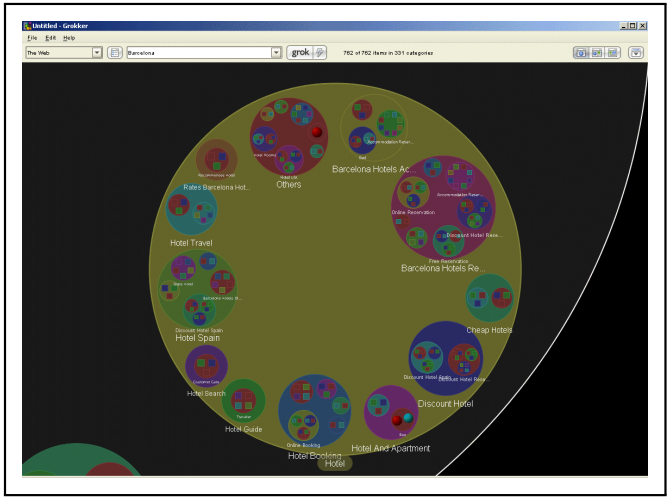
grokker (2)

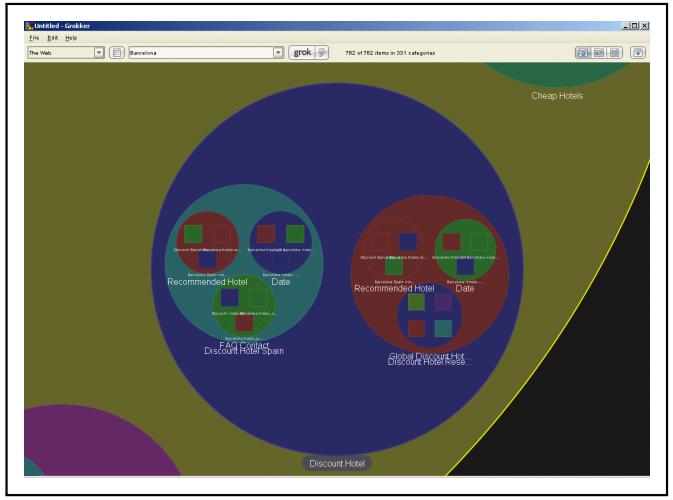

Grokker - Enterprise Search Management - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://live.grokker.com/grokker.html?query=advas%20Linux&Yahoo=true&numResul
grokker - Google-Suche
Grokker - Enterprise Searc...
semanticpool. current thoughts i...
See how Grokker can help your business
News & Events | Blogs | Contact Us | Feedback | Help | Home
grokker ESM
Selected Sources [1 of 3] Add/Remove
Yahoo!
advas Linux
GROK
Search Options
Outline View
Map View
59 total results
Working List
0 items in your list
View your list
Email Outline...
Export Outline...
Search within the outline:
by keyword
Exclude
by date
all
most recent
by source
<All sources>
by domain
<All domains>
Hide Tools
Expand Outline | Collapse Outline
advas Linux (59 results)
General (42)
3ddesktop (6)
Kdebase Kdepim Bind9 (5)
Debian Pool (3)
News (3)
Detail: Less Medium More
advas Linux > General
SourceForge.net: Files
Add to Working List | Post to del.icio.us | Bookmark | Email
You have selected to download the advas-0.2.1 release. ... Linux.com. NewsForge. freshmeat. Newsletters. PriceGrabber. Jobs. Find a Tech Job ...
http://sourceforge.net/project/showfiles.php?group_id=60060&package... - 26k - 11.02.2007
Source: Yahoo!
RPM Search
Add to Working List | Post to del.icio.us | Bookmark | Email
RPM PBone Search ... usr/share/doc/advas-0.2.0/doc/BUGS
/usr/share/doc/advas-0.2.0/doc/CHANGELOG ...
usr/share/doc/advas-0.2.0/doc/html/de/Chapter1_1.html ...
http://rpm.pbone.net/index.php3/stat/6/idpl/1575808 - 20k - 25.03.2007
Source: Yahoo!
RPM of Group Development/Libraries
Add to Working List | Post to del.icio.us | Bookmark | Email
advas-0.2.2-1. Python library for advanced search. linux/noarch. advas-0.2.1-1 ...
advas-0.2.0-1. Library for advanced search. linux/noarch. aes-1.1-1 ...
http://rpmfind.net/linux/RPM/sourceforge/Development_Libraries.html - 525k - 30.03.2007
Source: Yahoo!
http://ftp-master.debian.org/~joerg/joey/08.txt
Add to Working List | Post to del.icio.us | Bookmark | Email
Read live.grokker.com

grokker (3)


Others
Apartments
Hotel
barcelona
barcelona
notícies
destaquem
agenda

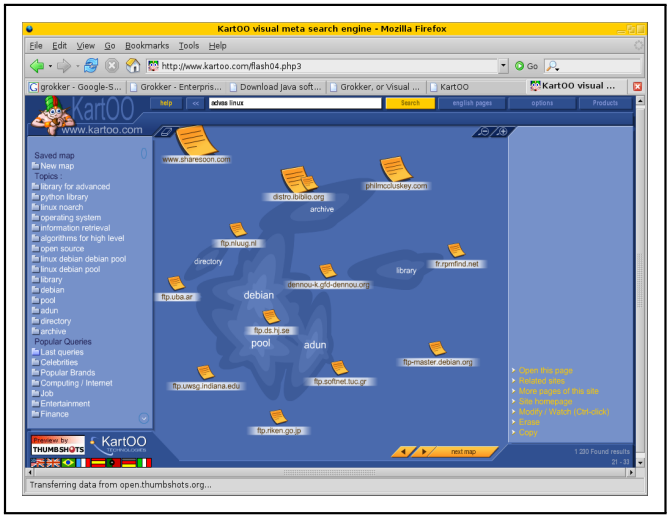

Others
Apartments
Barcelona's City
Catalunya
Guide
Hotel
Spanish
Travel
Web
Online
Barcelona


Others
Barcelona Hotels Ac
Hotel Travel
Barcelona Hotels Re
Cheap Hotels
Hotel Spain
Hotel Search
Hotel Guide
Discount Hotel
Hotel And Apartment
Hotel


Cheap Hotels
Recommended Hotel
Date
Discount Hotel Spain
Recommended Hotel
Date
Discount Hotel

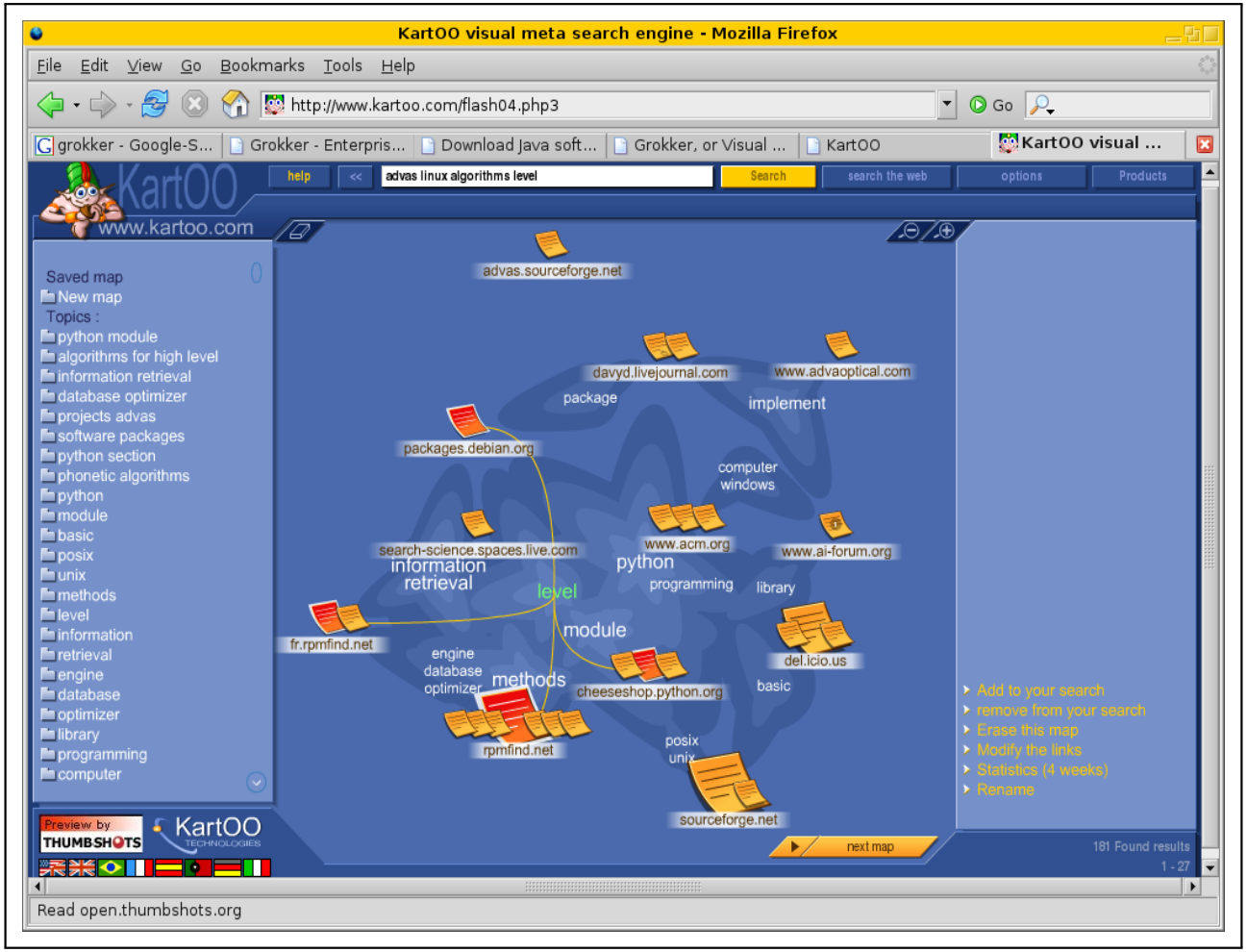
KartOO (1)

`http://www.kartoo.com`



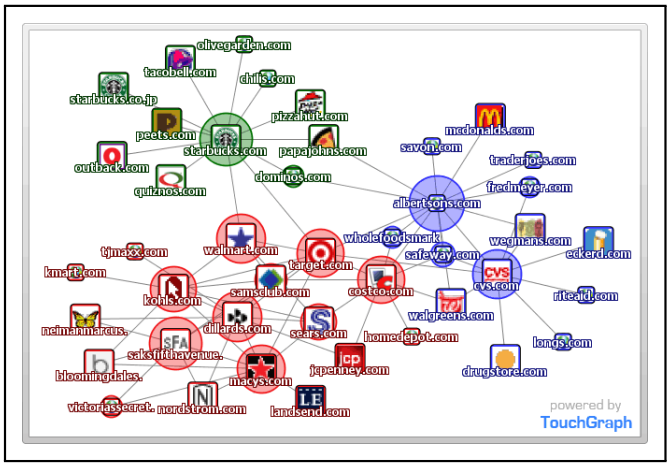
- Nutzung verschiedener Suchdienste
- graphische Darstellung auf der Basis von Flash
- Größe der Kreise entsprechend der jeweiligen Relevanz
- semantische Verbindungen zwischen den Kreisen

KartOO (2)


KartOO visual meta search engine - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.kartoo.com/flash04.php3
KartOO visual ...
www.kartoo.com
advas linux algorithms level
Search
search the web
options
Products
Saved map
New map
Topics :
python module
algorithms for high level
information retrieval
database optimizer
projects advas
software packages
python section
phonetic algorithms
python
module
basic
posix
unix
methods
level
information
retrieval
engine
database
optimizer
library
programming
computer
advas.sourceforge.net
davyd.livejournal.com
www.advaoptical.com
package
implement
packages.debian.org
computer
windows
search-science.spaces.live.com
information
retrieval
www.acm.org
python
programming
www.ai-forum.org
library
level
module
fr.rpmfind.net
engine
database
optimizer
methods
cheeseshop.python.org
basic
del.icio.us
rpmfind.net
posix
unix
sourceforge.net
Add to your search
remove from your search
Erase this map
Modify the links
Statistics (4 weeks)
Rename
Preview by
THUMBSHOTS
KartOO
next map
181 Found results
1 - 27
Read open.thumbshots.org

TouchGraph

`http://www.touchgraph.com`



- Nutzung verschiedener Suchdienste
- graphische Darstellung
- semantische Verbindungen zwischen den Kreisen

Literatur (Auswahl)

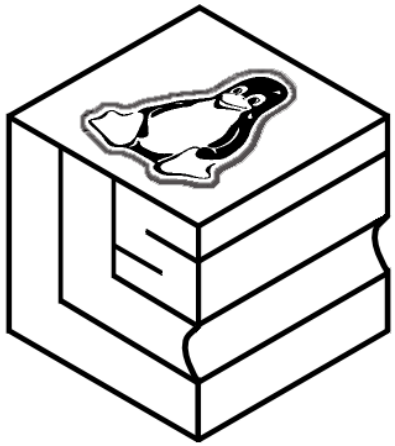
- G. G. Chowdhury:
  Introduction to Modern Information Retrieval
  Facet Publishing, London, 2004, 474 S., ISBN 1-8560-4480-7
- David A. Grossman/Ophir Frieder:
  Information Retrieval: Algorithms and Heuristics (The Information Retrieval Series), Springer, 2006, 356 S., ISBN 978-1402030048

Links

- Semantic Web School – Zentrum für Wissenstransfer
  `http://www.semantic-web.at`
- Semantic Pool
  `http://www.semanticpool.de`
- SearchEngine Watch
  `http://www.searchenginewatch.com`
- Suchtheorien und -strategien
  `http://wiki.apache.org/nutch/Search_Theory`

The End

Danke für Ihre Aufmerksamkeit :-)



Kontakt:
Dipl.-Inf. Frank Hofmann
Hofmann EDV – Linux, Layout und Satz
14467 Potsdam

Email <frank.hofmann@efho.de>
web www.efho.de