

Apache Lucene

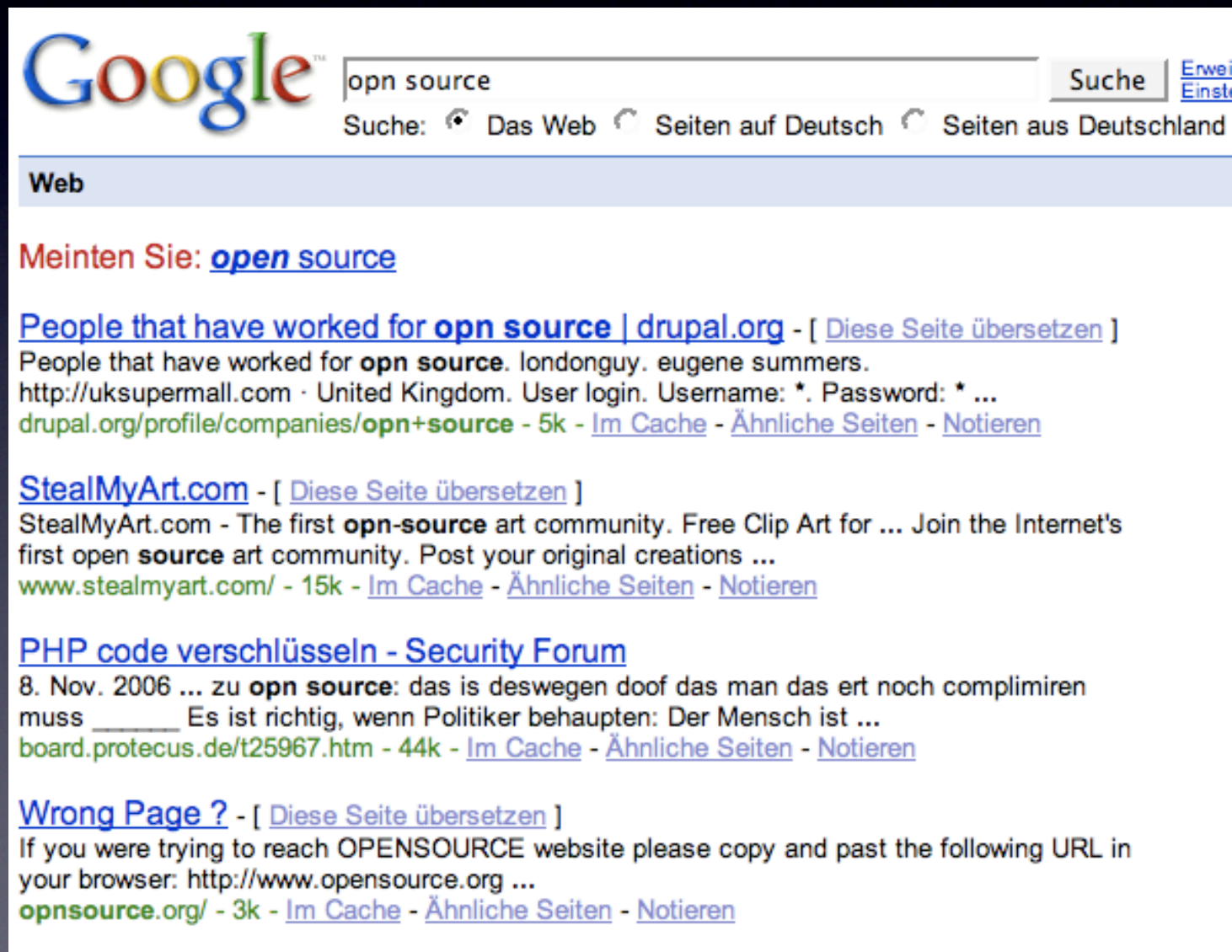
Mach's wie Google!

Bernd Fondermann
freier Software Architekt
bernd.fondermann@brainlounge.de
berndf@apache.org

Apache

- Apache Software Foundation
 - Software “free of charge”
 - Betonung liegt auf der Community
- Apache Software License
 - Free + Open + Source
 - problemlos in Closed Source verwendbar
- siehe ASF Session um 16:50 Uhr

Was macht Google, eigentlich?



The image shows a screenshot of a Google search result page. At the top left is the Google logo. To its right is a search bar containing the text 'open source'. Further right is a 'Suche' button and a link for 'Erweiterte Einstellungen'. Below the search bar, there are three radio buttons: 'Das Web' (selected), 'Seiten auf Deutsch', and 'Seiten aus Deutschland'. The main content area is titled 'Web' and lists several search results. Each result includes a title, a brief description, and a URL. The results are: 1. 'People that have worked for open source | drupal.org' with a link to translate the page. 2. 'StealMyArt.com' with a link to translate the page. 3. 'PHP code verschlüsseln - Security Forum' with a date and a link to translate the page. 4. 'Wrong Page ?' with a link to translate the page.

Google™ Suche [Erweiterte Einstellungen](#)

Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland

Web

Meinten Sie: [open source](#)

[People that have worked for open source | drupal.org](#) - [[Diese Seite übersetzen](#)]
People that have worked for **open source**. londonguy. eugene summers.
<http://uksupermall.com> · United Kingdom. User login. Username: *. Password: * ...
drupal.org/profile/companies/open+source - 5k - [Im Cache](#) - [Ähnliche Seiten](#) - [Notieren](#)

[StealMyArt.com](#) - [[Diese Seite übersetzen](#)]
StealMyArt.com - The first **open-source** art community. Free Clip Art for ... Join the Internet's first open **source** art community. Post your original creations ...
www.stealmyart.com/ - 15k - [Im Cache](#) - [Ähnliche Seiten](#) - [Notieren](#)

[PHP code verschlüsseln - Security Forum](#)
8. Nov. 2006 ... zu **open source**: das is deswegen doof das man das ert noch complimiren muss _____ Es ist richtig, wenn Politiker behaupten: Der Mensch ist ...
board.protecus.de/t25967.htm - 44k - [Im Cache](#) - [Ähnliche Seiten](#) - [Notieren](#)

[Wrong Page ?](#) - [[Diese Seite übersetzen](#)]
If you were trying to reach OPENSOURCE website please copy and past the following URL in your browser: <http://www.opensource.org> ...
opensource.org/ - 3k - [Im Cache](#) - [Ähnliche Seiten](#) - [Notieren](#)

ohne Freitextsuche

- Navigation in Bäumen
- Titel des gesuchten Dokumentes ist wesentlicher Anhaltspunkt
- öffnen vieler Dokumente, die nicht gesucht wurden
- zeitaufwendig
- alle Aktionen werden vom User ausgeführt
- Erwartungshaltung des Users: “googeln”

mit Freitextsuche

- alle Treffer auf einer Ebene, keine Bäume
- Treffer innerhalb von Dokumenten
- relevante Treffer sind schnell zugänglich
- Vorschau erlaubt direktes Filtern

Ablauf einer Freitextsuche

- vor der Suche (Computerprogramm)
 - Texte (Webseiten) finden und einlesen (Crawler)
 - 'Index' aufbauen
 - Index speichern und bereitstellen
 - Webseite mit Suchformular

Ablauf einer Freitextsuche

- während der Suche
 - Suchanfrage eingeben (User)
 - Stichworte entgegennehmen (Computer)
 - daraus eine Suchanfrage bilden
 - Index durchsuchen (schnell!)
 - Ergebnisse nach Relevanz sortieren
 - Ergebnis, inkl. Vorschau anzeigen

Lucene Produkte

| | |
|---------------|--------------------------------|
| Apache Lucene | Suchmaschinen-Framework, Index |
| Apache Nutch | Crawler + Indexer |
| Apache Solr | Suchmaschinen-Server |

Apache Lucene Features

- Indizieren von beliebigen Business-Daten
- Ablage in speziellen Files
- Formulierung von Suchanfragen
- Durchführung der Suche (Finden)
- Rückgabe der Ergebnisse, Ranking

Eigene Daten einspeisen

- Text, Text, Text
- HTML, XML
- PDF, Office-Formate
- Product {Name, Datum, Beschreibung, Attribute, Typ}
- File {Name, Content, Kind-Files}

Sucheanfragen

- ‘premium’, ““top five” ‘
- ‘premium special’, ‘premium OR special‘
- ‘premium AND -name:premium‘
- ‘pre*’, ‘pre~‘
- ‘datum:[20070101 TO 20081231]’

Ergebnisaufbereitung

- Ranking der Ergebnisse
- Fundstellen in wichtigen Feldern: 'boosten'
- Vorschau der Fundstellen in Fließtexten

Apache Nutch

- Crawler und Indexer
- durchsucht HTML Seiten automatisch
- extrahiert die relevanten Informationen
- folgt Links
- speist Ergebnisse in Lucene ein

Apache Solr

- Such-Server Komplettlösung
- Crawling, Indexing, Caching, Updating
- Administrations-Interface
- Optimierungen für High-Load
- Plug-in Architektur
- eigene Dokumente per XML einspeisen

Community & Support

- Lucene ist Quasi-Standard bei Such-APIs
- Support auf den öffentlichen Mailing-Listen
- sehr aktives Projekt
- derzeit kein kommerzieller Service
- Einfluß auf Projekt durch aktive Mitarbeit

Lucene Limits

- kein verteilter, replizierter Index
- keine “Meinten Sie?” Funktion
- keine Algorithmen zur schnellen Analyse großer Datenmengen

verwandte Apache Produkte

| | |
|---------------|-------------------------------|
| Apache Hadoop | verteiltes Filesystem |
| Apache Hadoop | Map/Reduce |
| Apache Mahout | Maschinenlernen |
| Apache Pig | Analyse großer Datenmengen |

Apache Hadoop

- verteiltes, redundantes Filesystem HDFS
 - verteilt über Rechner, Racks, Datacenter
 - beliebig viele Live-Kopien, kein Offline-Backup nötig
- große Datenmengen, effizient laden
- verteilte Datenbank HBase
- Map/Reduce: Distributed Computing

Apache Mahout

- maschine learning
- neues Projekt
- implementiert herausragende ML-
Algorithmen
- auf Basis von Map/Reduce
- experimentell = noch keine Integration in
Lucene

ML-Algorithmen

| | |
|----------------|--------------------------------------|
| “Meinten Sie?” | anhand der vielen Suchanfragen |
| Autocompletion | anhand der beliebtesten Suchanfragen |
| Naive Bayes | Spam-Detection |

Vielen Dank!

- <http://lucene.apache.org>
- <http://hadoop.apache.org>
- <http://incubator.apache.org/pig>
- <http://apache.org>
- <http://lucene.apache.org/mahout>
- Fragen und Antworten