

Besser crawlen, schneller finden

Suchmaschinen
Im Allgemeinen und bei PANVISION

Felix Fürer
Björn Schmidt

Panvision November 2013

Roadmap

- Entstehungsgeschichte Suchmaschinen
 - Erweiterung der Fähigkeiten
- Technik im Hintergrund
- Die PANVISION Suche
- Fallbeispiele
- Live Präsentation
- Ausblick & neue Features

Entstehungsgeschichte - Steinzeit

- 1990 Archie
 - FTP Dateidatenbank über Telnet
 - Erster "Robot", aber nur Dateinamen
- 1991 Jughead / Veronica
 - Gopher Suche über eigenes Programm
 - Auch interessant: CCSO Nameserver
- 1992 VLib (*Tim Berners-Lee*)
 - Statischer WWW Index

Archie Query Form



Search for:



Entstehungsgeschichte - Eisenzeit

- 1993 World Wide Web Wanderer
 - Erster WWW-Robot
 - Idee war eher, das Web zu vermessen
 - Brachte zeitweilig das „Internet“ down



- 1994 Excite

- Erster WWW-Robot mit **Ranking**
- Fiel der Dotcom-Blase zum Opfer



Entstehungsgeschichte - Dotcom

- 1994 Yahoo

- Hand gepflegter Index
- Erstes kommerziell erfolgreiches Anzeigenmodell



- 1994 Altavista

- Suche in natürlicher Sprache
- Suchvorschläge
- Kennt es noch jemand?



Entstehungsgeschichte - Dotcom

- 1995 Lycos, Looksmart
 - Pay per click



Galaxy, Infoseek, WebCrawler

- Es wird exponentiell

- 1996 HotBot



- **Ernsthafter** Altavista / Google Konkurrent
- Übernahme durch Lycos dann Yahoo

Entstehungsgeschichte - Dotcom

- 1997 Ask

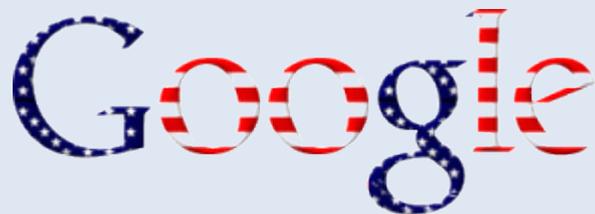


- Web Communities
- Erst Idee eines User-Ranking



- 1998 Google

- Don't be evil!
- Oder doch?



Entstehungsgeschichte - Neuzeit

- 1999 MSN, Alltheweb, dmoz, overture
 - Alle noch in Betrieb, immerhin
- 2005 2800 Days Later...
 - Alles entweder von Yahoo aufgekauft und zu Grunde gerichtet
 - Oder von Google assimiliert.
- 2008, 2009
 - Liveseach / Bing

Spezialisierte Suchmaschinen

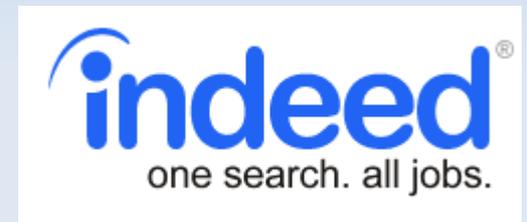
- 2002 Technorati

- Blog Suche



- 2004 Indeed Suche

- Evtl. auch interessant ;-)



- 1998 CiteSeer

- Wissenschaftliche Artikel

- 2008 deepdyve

- Wissenschaftliche Artikel (kommerziell)

Deep Web

Warum es die PANVISION Suche gibt

- Opaque Web / Private Web
 - Suchtiefe, Zugangsbeschränkung
- Invisible Web
 - Datenbanken ohne REST
- Truly Invisible Web
 - Javascript Navigation, Skripte
 - Alte Dokumentenformate

Anforderungen PANVISION Suche

- Durchsuchung von Deep Web
- Keine spezielle Programmierung
- Abbildung von Rollenkonzepten
- Kostengünstige Lösung

Dazu später mehr

Aber zunächst ein wenig Technik

Technik – damals und heute

- Präsentation anno 2000:
 - „Unglaublichen Datenmengen bei AltaVista“
 - 55 Millionen Seiten (heute eher 55 Trillionen)
 - 2 Millionen Zugriffe täglich
 - 16 Hochleistungs - Workstations der Reihe Alpha Server 8400 S/300
 - 6 Gigabyte Arbeitsspeicher
 - Häufig angefragte Stichwörter im Arbeitsspeicher

Abfrageanforderungen

- Einfache Wortsuche (Terms)
 - Ähnlichkeitssuche, Autovervollständigung
 - Rechtschreibung
- Wertebereichssuche
 - Datum, Werte
- Phrasensuche
 - Ähnlichkeit von verketteten „Terms“
- Verknüpfungen und Boosting

Technik im Hintergrund

- Robot / Spider
 - Durchsucht das Web (z.B. DFS, BFS, Ants)
 - Übergibt Gefundenes an den Indexer
- Analyzer / Tokenizer / Extraktor
 - Zerlegt die gelieferten Daten und organisiert sie
- Datenbank
 - Heutzutage Verteilt, Cached, Mehrstufig
- Abfrageserver

Analyzer / Tokenizer

- Rohdaten werden zerlegt
 - Metadaten (Keywords, Autor, Zeit, Ort,...)
 - Gewichteter Inhalt als Text
 - Text zugehörig zu Indexfeldern
 - Data Boosting für Metadaten und Text
- Daten werden in Terme zerlegt
 - Stopwords, Stopzeichen
 - Spracherkennung / Schriftzeichen

Datenbank – Invertierter Index

- Term Dictionary und Field Index
 - Index über alle Terme aus allen Dokumenten
 - Index über Terme in Feldern
- Term Vector, Proximity- und Frequenzdaten
 - In welchem Feld kommt der Term wie oft vor
 - In welchem Feld kommt der Term wo vor

Ablauf des Suchvorgangs

- Query Builder
 - Phrasen, Verknüpfungen, Boosting
- HitCollector Scorer
 - Finde Treffer im Verteilten Index
 - Bewerte Treffer z.B. über Boosting
- Filter und Sorter
 - Erstelle Ranking aus den Scorerwerten

Effizienzsteigerung

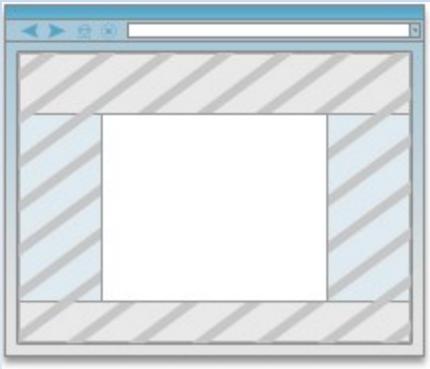
- Trigramme
 - Pan & anv & nvi & vis & Suc & uch & che
 - Für Autocomplete / Autosuggest
- Abfrageexpansion
 - Integriert Inhalte gefundener Dokumente in die Suche
- Abfragecache / Keywordcache
 - High Level als Proxy; low Level z.B. im HitCollector

Doch nun zur PANVISION Suche

- Identifizierung die zu durchsuchenden Datenquellen
 - Intranet (Private oder **Invisible Web**)
 - Dokumenten/Asset Management System (**Truly Invisible Web**)
 - Datenbank Applikation
 - Jobbörse
 - Telefonbuch

Vorbereitung der HTML-Daten

- Klassifizierung des Contents



- Über Metainformation
- Verstecken von Content
- Aufteilung in Blöcke

```
<!-- nxid=1 -->
<body>
  <div id="container">
    <div id="head">
      <p class>....</p>
    </div>
    <div id="sidebarLeft">
      <p class>....</p>
    </div>
  <!-- nxid=1 -->
    <div id="mainContent">
  <!-- contentid=1 -->
    <p class>....</p>
  <!-- contentid=1 -->
  <!-- contentid=2 -->
    <p class>....</p>
  <!-- contentid=2 -->
    </div>
  <!-- nxid=1 -->
    <div id="sidebarRight">
      <p class>....</p>
    </div>
    <div id="foot">
      <p class>....</p>
    </div>
  </div>
</body>
<!-- nxid=1 -->
```

Dublin Core Zusatzinfos

- ISAPI Filter oder IIS-Module HttpHandler
- Einbindung zusätzlicher Informationen per XSLT
 - Zugriffsberechtigte ID
 - Content-Typ
 - Lokalisierungs- / Sprachinfo

Suchhelfer für Datenbanken

- Bereitstellung einer Sitemap
 - Beispiel Telefonbuch
 - HTML Repräsentation einzelner Ressourcen

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
  <meta id="PVdetailPage" name="PV.detailpage" content=".../BusinessCard/id/208" />
  <meta id="PVthumb" name="PV.thumb" content=".../BusinessCard/userImage.ashx/208?maxHeight=170" />
  <meta name="DC.type" content="BusinessCard" />
  <meta name="PV.rating" content="0" scheme="PVTERMS.RATING" />
  <title>Mustermann, Max</title>
</head>
<body>
  <span id="detail">
    <br><!-- contentid=1 -->Mustermann, Max<!-- contentid=1 -->
    <br><!-- contentid=2 -->Max<!-- contentid=2 -->
    <br><!-- contentid=3 -->Mustermann<!-- contentid=3 -->
    <br><!-- contentid=4 -->musterman.max@company.com<!-- contentid=4 -->
    <br><!-- contentid=5 -->Specialist Corporate Communications<!-- contentid=5 -->
    <br><!-- contentid=6 -->mustermann<!-- contentid=6 -->
    <br><!-- contentid=7 -->+49 1234 123 9876<!-- contentid=7 -->
    <br><!-- contentid=8 -->+49 1234 123 1111<!-- contentid=8 -->
    <br><!-- contentid=9 -->9876<!-- contentid=9 -->
  </span>
</body>
</html>
```

Spider / Crawler Konfiguration

- Konfigurationsdatei bearbeiten
 - Einstellung der HTTP Startadressen
 - Hinterlegung der Zugangsdaten
 - Konfiguration der Filterung nach Dateitypen

Einrichtung des Frontends

- Service-orientierte Architektur (SOA)
- Proxy-Einrichtung
für den Abfrageserver
- HTML Suchformular
 - Einbindung der JS-Lib
 - Konfigurations-
Manuskript

```
var config = {
  'L10N' : {
    // lokalisierung der angezeigten Texte
  },
  'Server': '.../restproxy.php/search/infoservice/',
  '----': '...',
  ...,
  ,
  'Parts':
  [
    {
      'Header' : 'Intranet'
      // Suchfilter
      ...
    },
    {
      'Header' : 'Telefonbuch'
      // Suchfilter
      ...
    },
    {
      'Header' : 'Dokumente'
      // Suchfilter
      ...
    }
  ]
}
```

Fallbeispiele



Unternehmen Newsroom Karriere

Länderauswahl    Rolladen 

PRIVATKUNDEN

FACHPARTNER & ARCHITEKTEN

PRODUKTE

SERVICES

KONTAKT

Suchergebnisse

Suchergebnisse

15 hits

|« «« 1 2 »» »|

[890274_x.pdf](#)

... Rollläden an der **Rolladen** fährt nach Betätigung des örtlichen Tasters in die entsprechende Richtung und geht sofort in Selbsthaltung. Der Taster kann dann losgelassen werden und der **Rolladen** fährt bis zum...

[Windwarnanlagen_300806.pdf](#)

... Tabellen der Technischen Richtlinie des Bundesverbandes **Rolladen** Sonnenschutz e.V. Blatt 6.2 Seite 1 und... Richtlinie des Bundesverbandes **Rolladen** Sonnenschutz e.V. Blatt 6.2 Seite 1 und 2 festgehalten. Siehe Anhang...

[Presse_GHC.pdf](#)

... in Europa jetzt auf der internationalen Fachmesse für **Rolladen** Tore und Sonnenschutz R.T...

[EWFS Wandsender](#)

... PL FZL EWFS Uniswitch Integrierter Sensorik Solar **Rolladen** Integrierter Windsensor ISE...

[EWFS Handsender 1K](#)

... Integrierter Sensorik Solar **Rolladen** Integrierter Windsensor ISE Komfortsteuerung Comfort Timer...

Ausblick & neue Features

- AngularJS Client
 - MVVM Architektur
 - Anpassung, Lokalisierung
- Explanations
- Detailansicht-Index
- Adminhilfe-Menu
- Manuelle Indizierung

The screenshot displays a web application interface for 'Marketing'. The main heading is 'Marketing', followed by the subtitle 'Alles rund ums und aus dem Marketing'. The interface features a vertical navigation menu on the left side, organized into several sections:

- Formulare**: A folder icon with a minus sign (-).
- Corporate Design**: A folder icon with a minus sign (-).
- Bildrechte**: A folder icon with a plus sign (+), containing the text 'Einwilligungserklärungen für die verschiedenen Medien und Fotoerlaubnis'.
- Meldung Fotoobjekt.doc**: A document icon.
- Navigationssseite Corporate Design**: A document icon, with a sub-item 'Meldung Fotosafari'.
- Einkauf & Produktion**: A folder icon with a plus sign (+).
- KC Marketing**: A folder icon with a plus sign (+).
- Marktforschung**: A folder icon with a plus sign (+).
- Messe- und Ausstellungsbau**: A folder icon with a plus sign (+).
- Partner Servicecenter**: A folder icon with a plus sign (+).
- Werbelager**: A folder icon with a plus sign (+).
- Navigationssseite Formulare**: A document icon, with the text 'Hier finden Sie Formulare aus dem Marketing, geordnet nach den'.
- Information**: A folder icon with a plus sign (+).

On the right side of the interface, there is a context menu with the following options: 'Indiziere', 'Details', 'Rating', and 'Synonym'. Each option is accompanied by a magnifying glass icon and a plus sign with a bar chart icon. The overall layout is clean and professional, with a light blue and white color scheme.