Emeritenstamm, Winterthur, 30. März 2009

Warum ist Google so schnell?

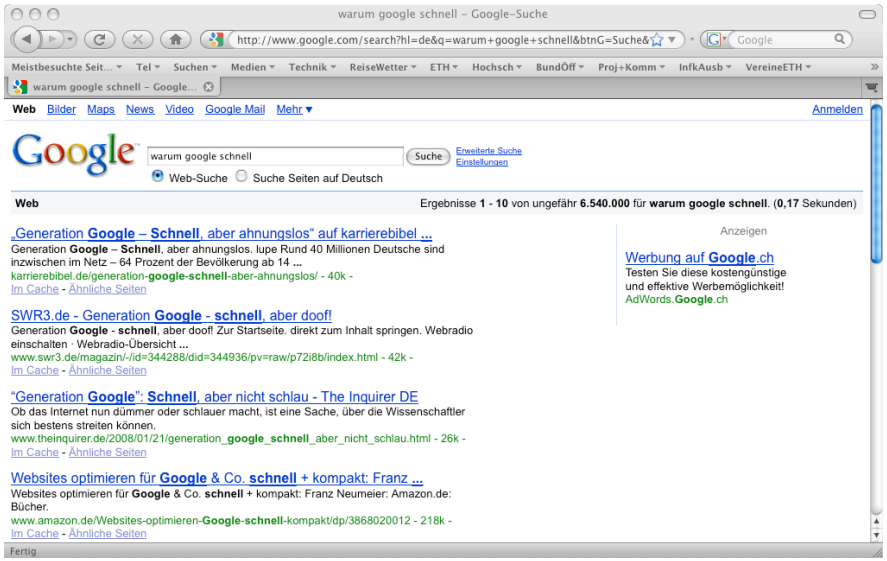
Carl August Zehnder
emeritierter Professor für Informatik ETH Zürich

© C.A. Zehnder, ETH Zürich. 2009, ergänzt 2010

Inhalt

- Warum? drei Antworten
- Informatik: Suchprozesse
- Internet und WWW
- Die Suchmaschine Google
- Schnell suchen mit Google
- Datenzentren
- Das Geschäftsmodell
- Kommentare

Warum ist Google so schnell? (1) **Die Antwort von Google**



Warum ist Google so schnell? (2)

Die Antwort der Informatiker

Google kombiniert alle Möglichkeiten der Informatik (= Informationstechnik) optimal und dies bei

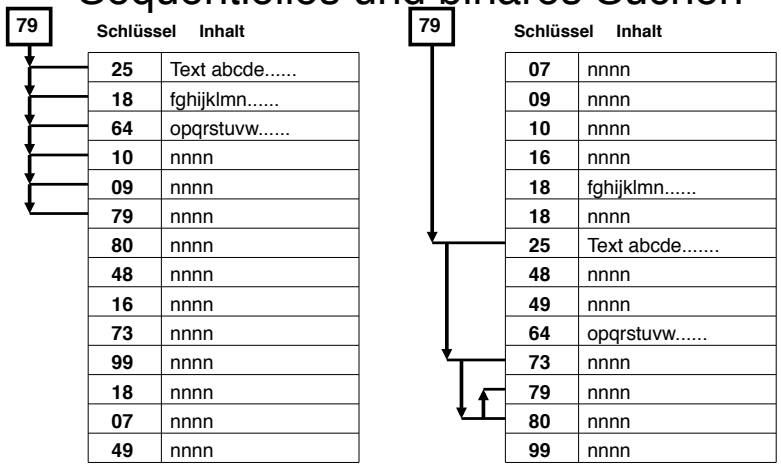
- Geräten (Hardware)
- Programmen (Software 1) und
- Daten (Software 2)

Die Antwort der Ökonomen

Das Geschäftsmodell von Google passt optimal in unsere gierige Zeit und macht so die ganze Welt zu begeisterten Kunden:

- Einfache Suchfragen akzeptieren.
- Raschmöglichst viele Antworten liefern.
- Lieber extrem schnell als "gut verdaut".
- Und das Ganze gratis (d.h. zu Lasten Dritter, die für Beachtung bezahlen).

Sequentielles und binäres Suchen



unsortierte Tabelle:
14 Datensätze, 6 Suchschritte

sortierte Tabelle
14 Datensätze, 4 Suchschritte

Suchaufwand in grossen Tabellen

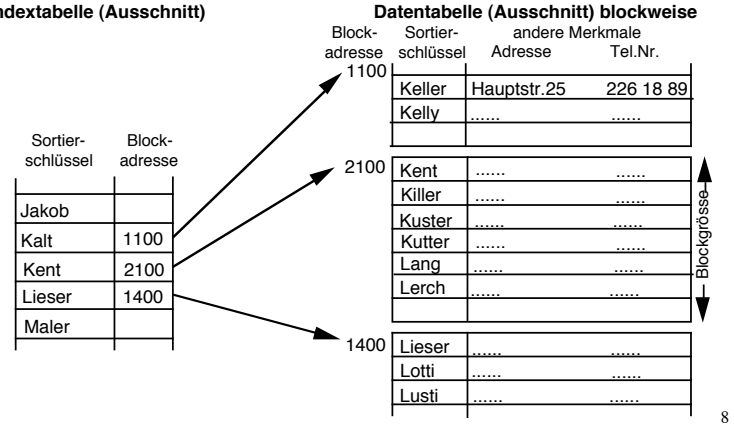
Beim binären Suchen steigt der Suchaufwand nur logarithmisch mit der Anzahl Datensätze.

		Anzahl Datensätze	**Anzahl Suchschritte sequentiell** (im Durchschnitt)	**Anzahl Suchschritte binär**
k	= Kilo	1000	500	10
M	= Mega	1'000'000	500'000	20
G	= Giga	1'000'000'000	500'000'000	30
T	= Tera	1'000'000'000'000	500'000'000'000	40
P	= Peta			
		n	**n/2**	**$_2$log n**

Hilfe beim Kopfrechnen: 2^{10} = 1'024, das ist ungefähr 1'000.

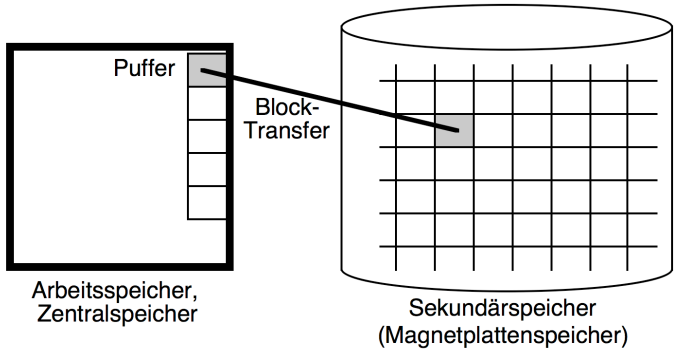
Indextabellen, indexsequentielle Org.

Grosse Tabellen werden in Datenblöcke aufgeteilt, die über eine Indextabelle (eine Art Inhaltsverzeichnis) direkt aufgerufen werden können.



Schnelle vs. grosse Speichermedien

Im Computer arbeiten Prozessor und Arbeitsspeicher im Nanosekundenbereich; der Zugriff auf den Plattenspeicher (der sog. Blockzugriff) ist sehr viel langsamer.



Zugriffszeit 0.1 - 1 μs
Kapazität viele MByte

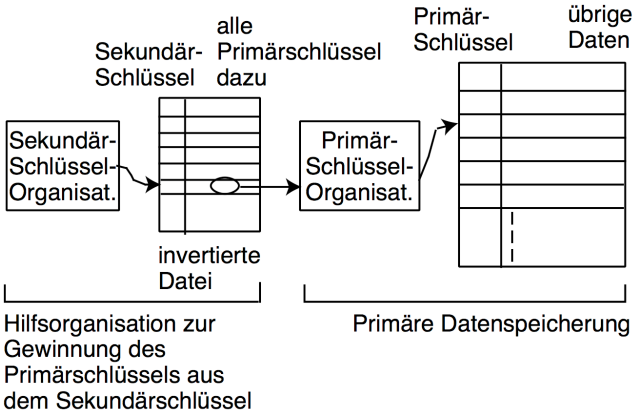
Zugriffszeit ca. 10 ms
Kapazität GByte - TByte

Beispiel Telefonbuch

- Telefonbuch Stadt Zürich: 400'000 Abonnenten
- Einstufiges Suchen binär:
  19 Suchschritte (2^{19} = ca. 500'000)
- Zweistufiges Suchen binär:
  1'000 Seiten (= Datenblöcke) mit je ca. 500 Datensätzen:
  In der Indextabelle für 1'000 Datenblöcke werden 10 Suchschritte (2^{10} = 1024) benötigt,
  in einem Datenblock 9 Suchschritte (2^9 = 512),
  total 10+9 = 19 Suchschritte.
  Das braucht nur zwei langsame Blockzugriffe!

Sekundärschlüssel

Schnelles Suchen ist nicht nur nach dem (sortierten) Primärschlüssel möglich; jeder Sekundärschlüssel benötigt aber entsprechende Hilfs-Indextabellen.



Beispiel ohne Computer: Spezialkataloge in grossen Bibliotheken

Schnelligkeit durch Parallelisierung

Die noch immer andauernde Leistungssteigerung der Informatik (Moore'sches Gesetz) beruht primär nicht auf schnelleren Prozessoren, sondern auf *viel mehr* Prozessoren und deren parallelem Einsatz. Stichworte dazu:

- Multi-Core-Systeme: Viele Prozessoren teilen sich in eine Arbeit.
- Cloud Computing: Alle verfügbaren Geräte (Hardware-Ressourcen: Speicher, Prozessoren, Netzwerke) werden als virtuelle Gesamtheit gesehen und flexibel Aufgaben zugeordnet.
- Programmierung paralleler Prozesse.

Verwandte Informatik-Aufgaben

- **Sortieren:** Voraussetzung jeder binären Suche sind nach dem Schlüssel vorsortierte Tabellen, Sortieren ist bei grossen Tabellen aufwendig (proportional zu n*log n).
- **Mutieren:** Schwieriger als das Abfragen (Suchen) ist in grossen Datentabellen das effiziente und sichere Nachführen der Daten (Ändern und Ergänzen). Indexsequentielle Tabellen lassen sich sehr effizient mutieren.

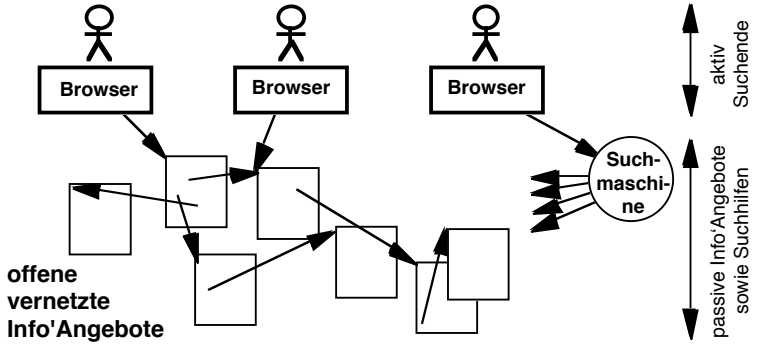
Sortieren und Mutieren sind zwei höchst interessante Aufgaben der Informatik, können hier aber nicht weiter behandelt werden.

Das Internet

- Das Internet wurde als Netzverbund ab 1969 (Arpanet) entwickelt, wichtige Beiträge stammen von Vincent Cerf und Robert Kahn.
- Die Internet-Kernkonzepte sind ein globales, dezentral verwaltetes Adresssystem für Dokumente (name.ch) und klar geregelte Schnittstellen (Protokolle).
- Das Internet wurde lange vor allem im Hochschulbereich genutzt. Wichtigste frühe Dienste: E-Mail, File Transfer, Newsgroups.
- Erst die Erfindung des WWW (Tim Berners-Lee, 1989-93, Cern, Genf) machte das Internet für jedermann nutzbar.

Adressierte Dokumente + Hyperlinks

Alle Dokumente sind adressiert. Wer ihre Adresse kennt, kann sie mit einem Browser im WWW aufrufen. Graphische Browser sind ab 1993 verfügbar.



Milliarden von Dokumenten („Webseiten“)

Einstieg über Adresse im WWW

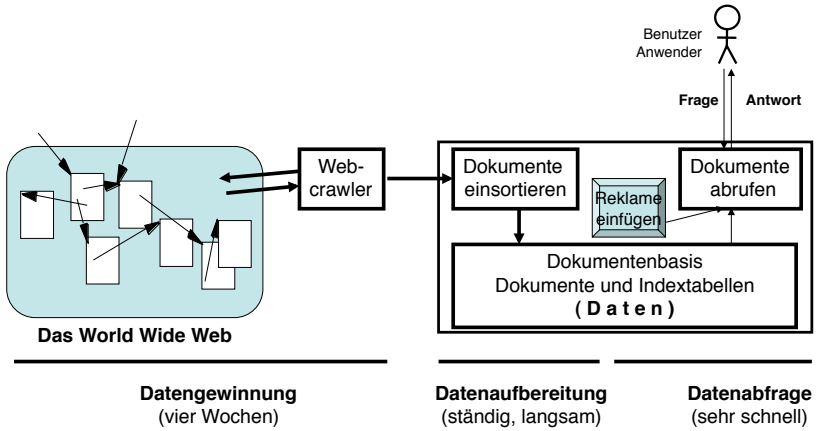
http://www.inf.ethz.ch/personal/zehnder/
www.google.com

Vom Präzisen zum Bequemen:

- Der Benutzer kennt die exakte Adresse.
- Der Benutzer kennt ein Portal, das ihm weiterhilft (mit Links, Navigationsbalken, Suchhilfen usw.)
- Der Benutzer kennt einen Katalog (Bsp. Wikipedia).
- Der Benutzer kennt die Adresse einer Suchmaschine (Bsp. Google).

Gesamtüberblick

Datengewinnung, -aufbereitung und -abfrage sind getrennt



Webcrawler

Webcrawler "durchkriechen" das ganze WWW regelmässig und besuchen alle zugänglichen Webseiten (Webpages):

- Neu gefundene Seiten werden in die Dokumentenbasis von Google kopiert.
- Geänderte Seiten werden in der Dokumentenbasis von Google ersetzt.
- Alle Seiten werden bewertet (Google: **PageRank™** Technology).

Für den PageRank werden etwa 200 Kriterien ausgewertet. Hoch bewertet werden namentlich Seiten, auf welche viele andere hoch bewertete Seiten Links gesetzt haben.

Ganz vorne auf die Google-Liste!

Der PageRank bestimmt die Reihenfolge der Google-Antworten. Diese ist für *alle* Beteiligten sehr wichtig:

- Viele Anwender schauen nur die vordersten Google-Antworten an.
- Die Web-Autoren wollen, dass ihre Seite rasch gefunden wird und möglichst unten den ersten zehn Google-Antworten erscheint.
- Google ist interessiert, dass ihre Rangliste den Anwendern bestmöglich dient (denn nur dann bleibt Google die Nr. 1 bei den Suchmaschinen im WWW).

Die Berechnung des PageRanks bildet die Kernaufgabe im Google-Prozess. Dieser Algorithmus ist geheim und wird nach Bedarf angepasst.

Finden Webcrawler alles?

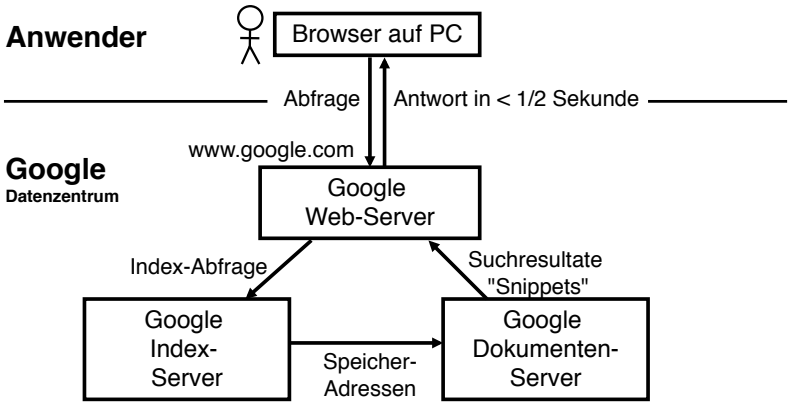
Nein!

- Webcrawler arbeiten ausschliesslich mit gültigen Web-Adressen aus Links in aktuellen, funktionierenden Web-Seiten. (Daher Vorsicht mit alten, nicht nachgeführten, gebrochenen Links!)
- Eine Web-Seite, auf die kein Link führt, wird vom Webcrawler nicht gefunden und somit auch nicht durch Google referenziert.
- Jedermann kann "geheime" Seiten auf seinen Web-Server hochladen, die man nur findet, wenn man die genaue Web-Adresse bereits kennt.

Das sog. "Deep Web" mit den nicht (mehr) verlinkten Seiten ist übrigens umfangreicher als das verlinkte Web.

Google: Schnelle Abfragen (1)

Ablauf einer Google-Abfrage

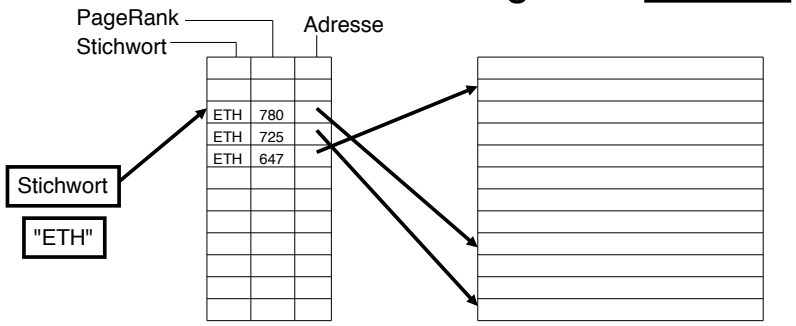


nach http://www.google.com/intl/en/corporate/tech.html
"Life of a Google Query"

Google: Schnelle Abfragen (2)

Ein-Wort-Abfragen

wenig präzis



Index-Server

enthält Verweise auf alle Web-Seiten, sortiert nach Stichworten und PageRank

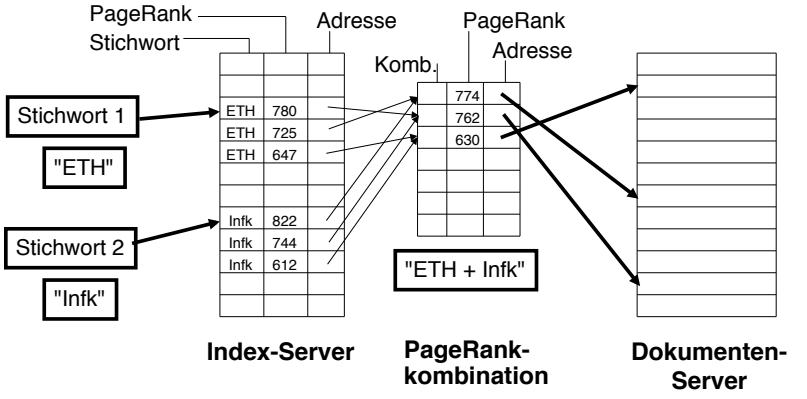
Dokumenten-Server

enthält die vollständigen Web-Seiten, nachgeführt durch den Webcrawler und sortiert nach interner Adresse

Google: Schnelle Abfragen (3)

Mehr-Wort-Abfragen

viel präziser!



Index-Server

PageRank-kombination

Hilfstabelle zur Berechnung kombinierte PageRanks

Dokumenten-Server

Google: Schnelle Abfragen (4)

Google ist unkompliziert

Google macht es seinen Kunden so einfach wie möglich, Abfragen ans WWW zu formulieren:

- Schon die Eingabe eines einzigen Stichwortes genügt als Abfrage.
- Häufig gebrauchte Stichwörter werden auch akzeptiert, wenn sie orthographisch falsch geschrieben werden. (Dahinter stehen leistungsfähige Hilfsprogramme und sog. "künstliche Intelligenz").
- Bei der Verwendung mehrerer Stichwörter genügt eine schlichte Aufzählung (Satzzeichen oder gar mathematische Spezialzeichen werden nicht benötigt).

Google: Schnelle Abfragen (5)

Kluge Anwender – bessere Antworten

"Googeln" steht als Begriff bereits im Duden; aber viele Anwender kennen folgende Regeln noch nicht:

- Breite Suchbegriffe (Bsp. "Radio") liefern wenig präzise Suchresultate und viel Nutzloses.
- Die Verwendung von Einzelwörtern als Suchbegriff ist unprofessionell. Wortkombinationen führen direkter zum Ziel (Bsp. "Radio Sender").
- Unerwünschte Begriffe können mit einem Minus direkt ausgeschlossen werden (Bsp. "-DRS").

Google: Datenzentren (nach Friedemann Mattern)

Google-Datenzentren weltweit

- Jedes Datenzentrum hat 10 000 – 100 000 Computer
- kostet über 500 Mio $ (Bau, Infrastruktur, Computer)
- verbraucht 50 – 100 MW Energie (Strom, Kühlung) **=SBB / 2**
- Neben Google weitere (z.B. Amazon, Microsoft, Ebay,…)

Google Data Center Columbia River

Rückansicht: Energiezufuhr

Innenansicht

- Effizient wie Fabriken
  - Produkt: Internet-Dienste
- Kostenvorteil durch Skaleneffekt
  - Faktor 5 – 7 gegenüber traditionellen „kleinen“ Rechenzentren
- Angebot nicht benötigter Leistung auf einem Spot-Markt

Das entwickelt sich zum eigentlichen Geschäft!

Zukünftige Container-Datenzentren

- Hunderte von Containern aus je einigen tausend Compute-Servern
  - mit Anschlüssen für Strom und Kühlung
- Nahe an Kraftwerken
  - Transport von Daten billiger als Strom

- Anmerkung zum Thema „Green IT“:
  - IT-Infrastruktur verursachte 2007 weltweit ca. 830 Mio t CO_2 (2%)
  - IT-gestütztes Energie- und Gebäudemanagement sollte aber bis zu 4 Mal so viel einsparen

Google: Our philosophy

Larry Page's far reaching vision:
"The perfect search engine would understand exactly what you mean and give back exactly what you want."

- Focus on the user and all else will follow.
- It's best to do one thing really, really well.
- Fast is better than slow.
- Democracy on the web works. *
- You can make money without doing evil. *
- The need for information crosses all borders.
- Great just isn't good enough.

* -> das stösst an Grenzen!

Erfolg motiviert

Google hat das WWW nicht erfunden, aber für dieses einen hervorragenden Zusatzdienst entwickelt. Und eine Weltfirma:

- Google holt sich die besten Informatik-Ingenieure.
- Google bietet hervorragende Arbeitsbedingungen (selbständige Arbeit und viele Freiheiten)
- Google setzt hohe und laufend neue technische Ziele.
- Google setzt hohe ethische Standards und stellt sich auch heiklen Diskussionen (Bsp. Datenschutz bei Personendaten)
- Google verdient mit seiner Marktposition sehr viel Geld (die eingeblendeten Werbebegriffe werden laufend an den Meistbietenden versteigert).

Schlusskommentare (1)

Die Schwächen des Web sind auch die Schwächen von Google

Google liefert schnell, aber mit beschränkter Qualität.

- Im WWW kann jedermann Beliebiges aufschalten; viele dieser Informationsquellen sind daher unsicher. Google basiert somit auf unsicheren Quellen.
- Google will unbedingt schnell sein. Dessen Informationsbewertungsalgorithmen (PageRank) sind hochraffiniert, bringen jedoch nicht zwingend die besten Web-Dokumente auf die ersten Positionen.
- Viele Anwender von Google nehmen sich zu wenig Zeit, die Qualität der Suchergebnisse zu überprüfen.

Schlusskommentare (2)

Google passt in unsere Zeit

- Google nutzt die verfügbare Technik hervorragend.
- Google trifft die Wünsche der meisten Anwender besser als jede Konkurrenz.
- Google lässt sich von jenen entschädigen, die Kunden anwerben wollen; Werber bezahlen gerne für gezielte "Aufmerksamkeit" nahe bei Fragen der Kunden, und zwar pro Klick!

Schlusskommentare (3)

Auch Google stösst an Grenzen

- **Google Earth:**
  "Alles fotografieren und öffentlich anbieten!"
  -> verletzt Persönlichkeitsrechte
  -> Datenschutzbeauftragter verbietet Aufnahmen
  -> Versuche mit Anonymisierung (noch ungenügend)
- **Google Buchsuche (e-Library):**
  "Alle Bücher einscannen" und öffentlich anbieten!"
  -> verletzt Urheberrechte
  -> Klagen
  -> Verhandeln und Vergleich

>>> Unsere Rechtsordnung setzt Grenzen, damit die Rechte anderer geschützt werden können.

Schlusskommentare (4) – insbesondere für Studierende:

Überlegungen am TecDay

- Am Anfang (1998) stand eine Glanzidee.
  (Google: gratis schnell im WWW plus etwas Reklame)
- Diese Glanzidee nutzte aktuellste Technik.
  (Google: WWW, verfügbar ab 1993)
- Diese Glanzidee wird mit sehr guten Ingenieuren technisch ständig weiterentwickelt.
  (Google: Google Earth, e-Library usw.)
- Ein gutes Geschäftsmodell sichert Erfolg und Nachhaltigkeit des Unternehmens.
- Aber immer sind die Gesetze zu beachten.
  (sonst klagen Betroffene und Googles Konkurrenten)

>>> Gute Ausbildung als Basis aufbauen, Augen offen halten, Chancen erkennen und nutzen

Weiterführende Links

zu Suchprozessen (Informatik):

- http://de.wikipedia.org/wiki/Suchfunktion
- http://de.wikipedia.org/wiki/Suchmaschine
- http://www.google.com/intl/en/corporate/tech.html

zur Geschäftsphilosophie von Google:

- http://www.google.ch/intl/en/corporate/
- http://www.google.ch/intl/en/corporate/tenthings.html

zu Problementwicklungen von Google:

- http://de.wikipedia.org/wiki/Google_Street_View
- http://de.wikipedia.org/wiki/Google_Book_Search