Die Technik der Suchmaschinen

Wie durchsucht man

2,073,418,204 Web-Seiten

in 0.29 Sekunden?

Vortrag von Evelyne Kopf

Christof Kerkmann

Pascál Bihler

22. August 2002

IR – Information Retrieval

Informationssysteme

Datenbanken

* Relationale oder Objektorientierte Datenbanksysteme (MySQL, Oracle, SQLServer, Tamino...)
* Workflow-Systeme (SAP R3,...)
* Mangament-Informationssyteme

Dokumentationssysteme

* Literaturrecherche
* Dokumentation
* Archivierung
* Internet-Suchmaschinen

Ausgezeichnete Information

```
<Buch>
        <Titel>Durch die Nacht</Titel>
        <Autor>Heinz Müller</Autor>
        <ISBN>3-23415-4141-X</ISBN>
</Buch>
```

Vorteile:

* gute Durchsuchbarkeit
* berechenbare Suchergebnisse
* semantischer Gehalt bekannt

Nachteil:

* Daten müssen von Hand aufbereitet werden

Semantikarme Information

Neu bei uns verfügbar:

Müller, Heinz: **Durch die Nacht**

(3-23415-4141-X)

Vorteile:

* Daten können so übernommen werden, wie sie zur Verfügung stehen
* Keine Overhead durch Meta-Informationen

Nachteil:

* Schlechte Durchsuchbarkeit
* Fehlende semantische Information

Vollständige Invertierung

Neu bei uns verfügbar:

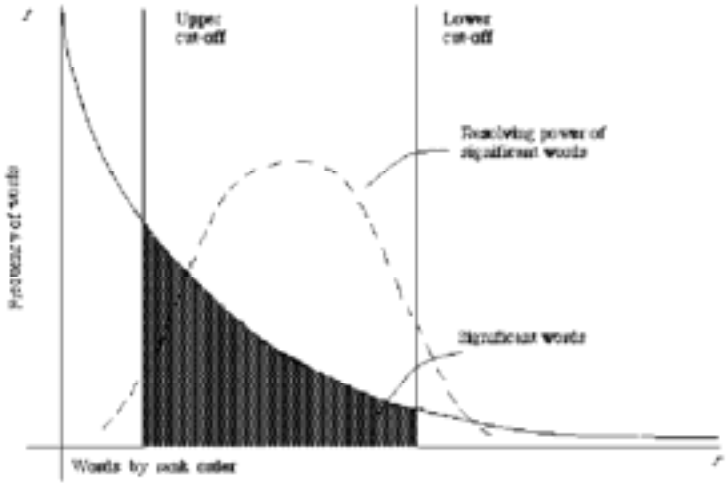
Müller, Heinz: **Durch die Nacht**

(3-23415-4141-X)

Vollständige Invertierung

```
BEI: 5
DIE: 46
DURCH: 40
HEINZ: 33
MÜLLER: 25
NACHT: 50
NEU: 0
UNS: 9
VERFUEGBAR: 13
```

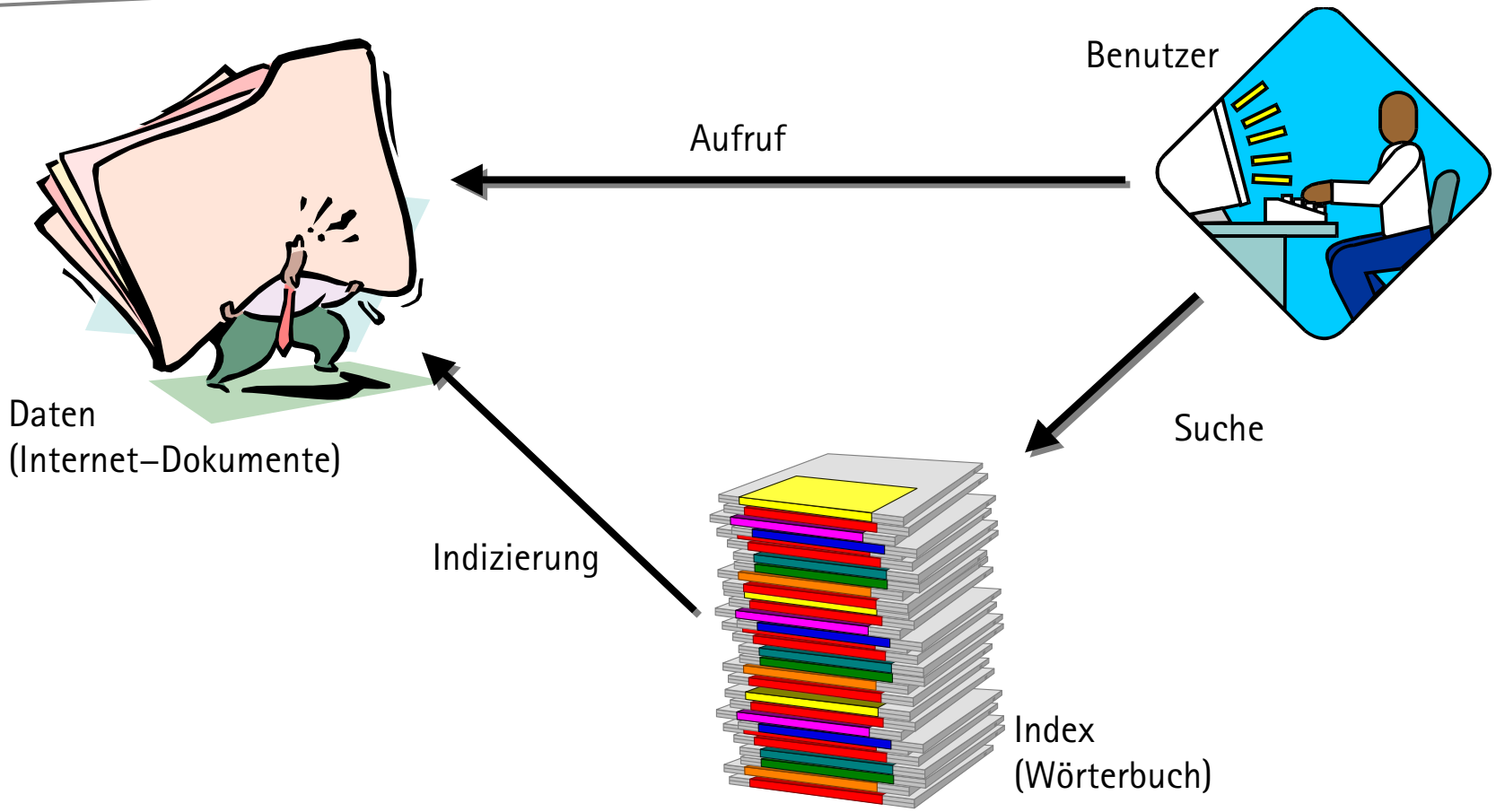
Stoppwörter



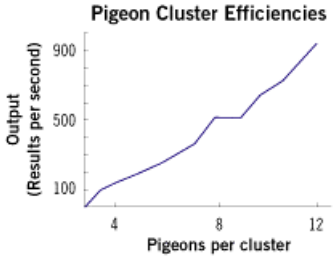
"To be or not to be"

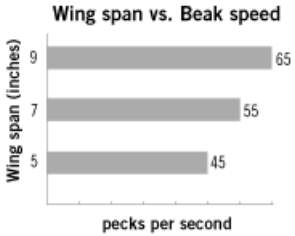
Englische Hochfrequenzwörter				
A	CANNOT	INTO	OR	THUS
ABOUT	CO	IS	OURS	TO
ABOVE	COULD	IT	OURSELVES	TOGETHER
ACROSS	DOWN	ITSELF	OUT	TOO
BE	DURING	NOT	OVER	TOWARD
...	...	...	...	...

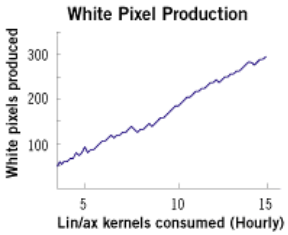
Allgemeines Prinzip der Indizierung

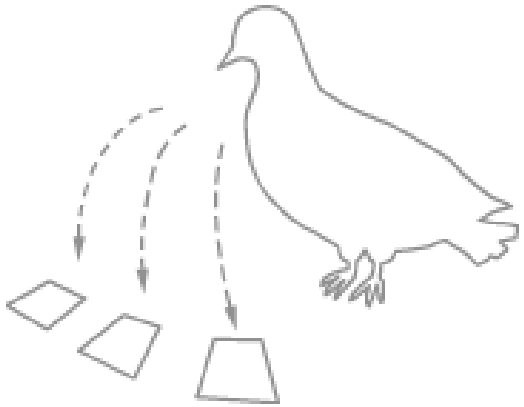


PigeonRank™



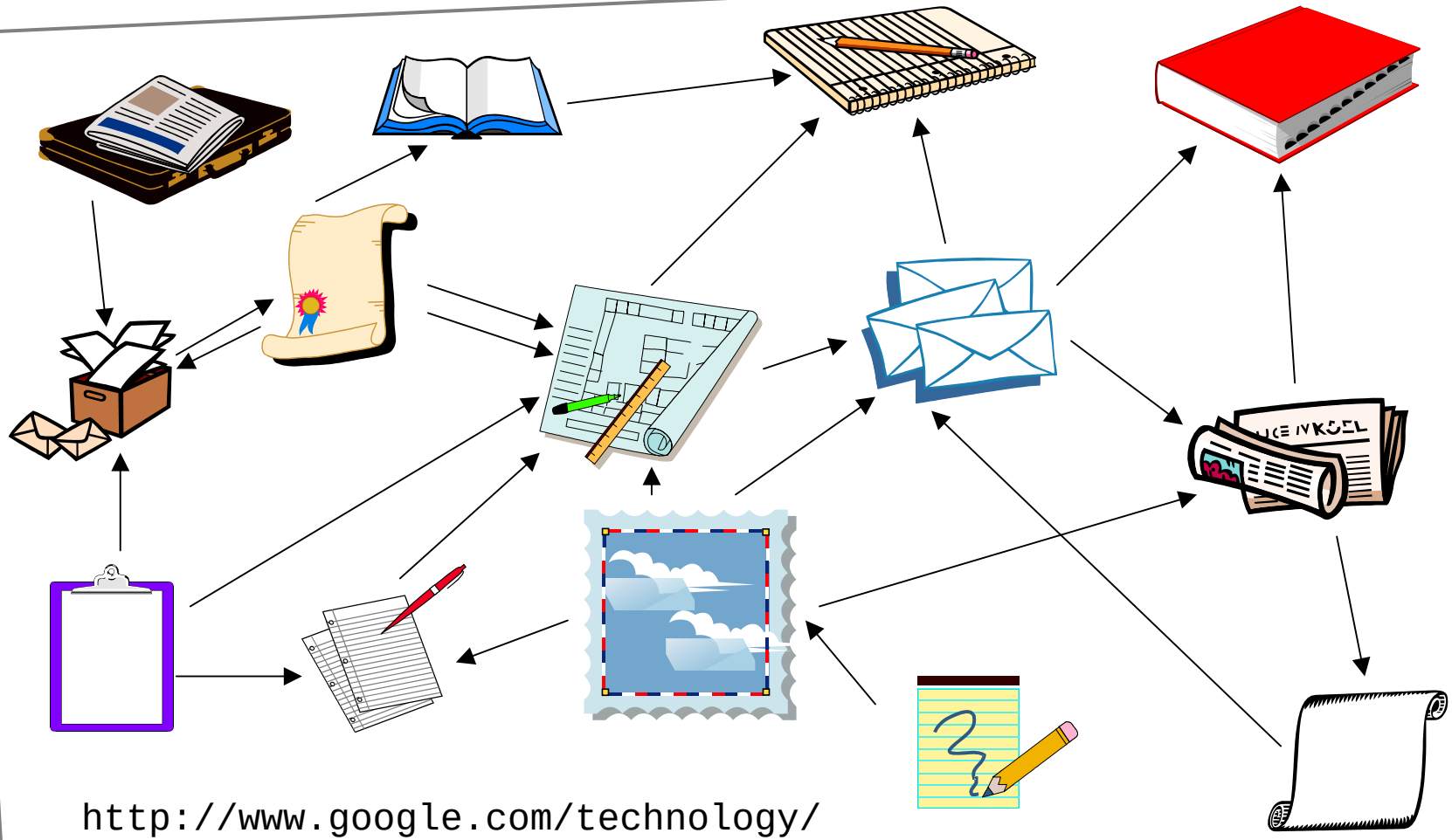






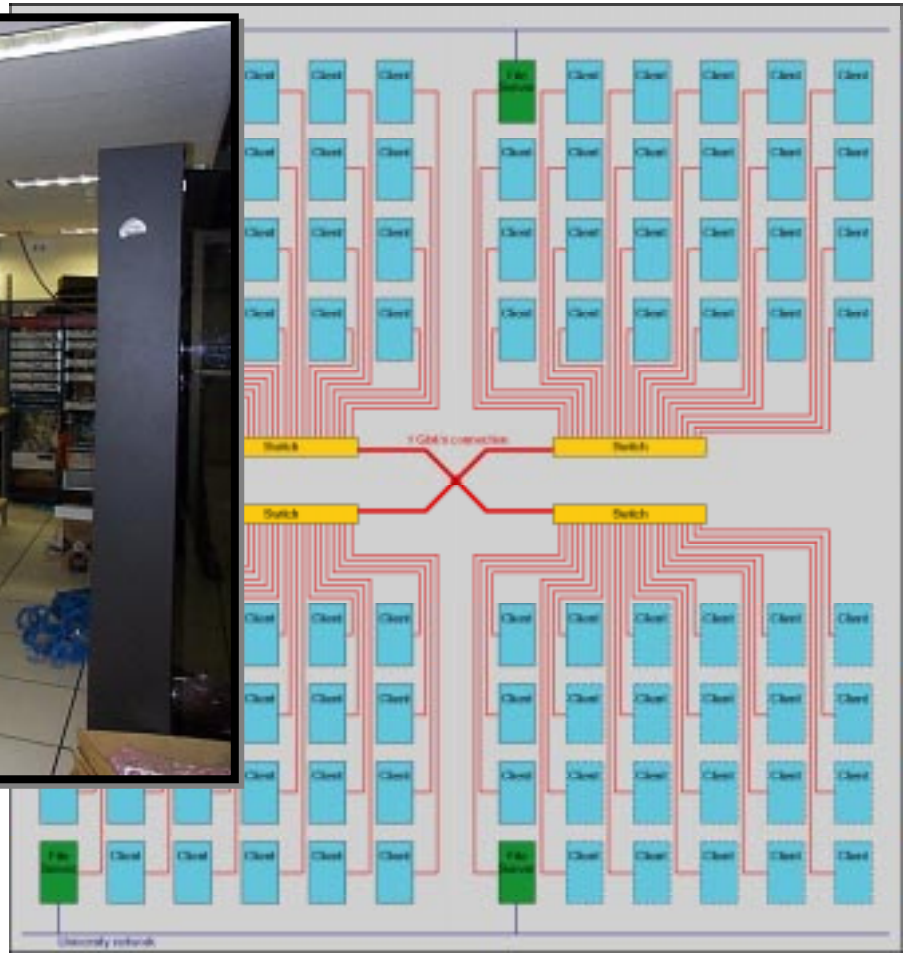
http://www.google.com/technology/pigeonrank.html

PageRank™



http://www.google.com/technology/
http://www-db.stanford.edu/pub/papers/google.pdf

Rechner-Cluster



Verweise

* Vorlesung Informatik IV – Uni Karlsruhe (TH) (A. Schmitt)
* Spektrum der Wissenschaft, 1999, Seite 44
* c't magazin für Computertechnik 07/02

* http://www.searchenginewatch.com/
* http://www.google.com/technology/
* http://www–db.stanford.edu/pub/papers/google.pdf
* http://kresch.com/search/searchterms.htm