Information Retrieval

Katja Markert

Institut für Computerlinguistik
Uni Heidelberg
markert@cl.uni-heidelberg.de

January 21, 2020

Bisher und heute

1. Bisher: Term-Term-Matrizen
2. Jetzt: Basis von Information Retrieval: Term-Document-Matrizen
3. Ausserdem: Inverted Indizes, TF-IDF etc
4. Ausserdem: Evaluation von Information Retrieval

Google Suche

panthers

Florida panthers official WebSite
. . . Panthers Sign Two, Release TV Schedule. Florida **P**anthers General Manager . . . **Panthers** Encores on Fox Sports Net. Tune into FOX Sports . . .
Description: Official site. Includes team information, statistics, a shedule, and ticket information.
Category: Sports > Hockey > . . . > Teams > Florida Panthers

Carolina panthers
. . . Weinke is having the training camp he was expected to have in 2002, leading the **Panthers** on all three of their touchdown drives in two

Gray panthers: Home
Gray **Panthers** National Office. 733 15th Street, NW Suite 437 Washington, DC 20005 (800) 280-2536 or (202(737-6637 Fax: (202) 727 . . .
Description: Advocacy group working on issues such as universal health care
Category: Regional > North America > . . . > Society and Culture > Seniors

Nottingham Panthers:
Panthers Official Hotline: 09068 800 660 (Calls cost 6p a minute at all times
PANTHERS AND NATIONAL ICE CENTRE REACH AGREEMENT FOR 2003/4 SEASON

Google Suche

panthers Africa

BBC SPORT — Football — Africa — When Panthers became Crocodiles
. . . will long remember the day when the Black Stars emerged 2-1 victors over the the Crocodiles who were supposed to be **Panthers**. . . . Links to more **Africa** stories are . . .
news.bbc.co.uk/sport2/hi/football/africa/2313895.stm - 59k - Cached - Similar pages

allAfrica.com: Zimbabwe: Panthers Upstage Busters to Remain in . . .
. . . **Panthers** Upstage Busters . . .
allafrica.com/stories/200306250562.html - Similar pages

allAfrica.com: Zimbabwe: Panthers, Oh Face Must-Win Ties
. . . Old Hararians vs MCD; Busters vs Mabvuku; Harare Sports Club vs Gweru; University vs Shabani; Western **Panthers** vs Old Miltonians; Old . . . **Africa** 2003 **Africa** 2003.

Africana.com: Gateway to the Black World
. . . Huey P. Newton, a founder of the Black **Panthers**, is shot and killed in . . . group of alleged white extremists facing treason charges in South **Africa** has complained . . .

NZOOM - ONE Sport - League
. . . **Panthers** shut-out Sharks. It was the ultimate payback and the ultimate irony. Sharks . . . berth. And scoring tries is the **Panthers**' specialty. . . .

Google Suche

panthers animals Africa

African & AustralAsian Animals
. . . **Africa** & AustralAsia, Elephants. Hippos & Rhinos. Kangaroos. Lions & Panthers.
Other. Europe & America. Polar & Maritime. Bears. Farmland. Forest. Pets. Flying **Animals**. . . .

Animals
. . . chase lions, **panthers**, leopards, and other large **animals** into an area surrounded
by shields and nets, such as is depicted in this mosaic from North **Africa**. . . .

ScreensaverShot.com - Animal Dogs screen savers(3) download and . . .
. . . Win98/Me/2000/XP; File Size : 2,564 KB; Description : Various images of **panthers**; . . . Me/2000/XP;
File Size : 2,009KB; Description : Various wild **animals** in **Africa**; . . .

ZOOS IN SOUTH AFRICA . . . the region of 2500 individual **animals** in South **Africa**. . . . The zoo has adult tigers, leopards,
panthers, a particularly . . . orangutan?s and many more **animals** in tiny . . .

BBC - Nature - Animals on the Edge: Cats . . . consisting of just 30-50 adult **animals** confined to . . . fences and underpasses
to give
panthers safe access . . . around protected areas in eastern and southern **Africa**. . . .

Information Retrieval: die Aufgabe

- **Gegeben:** große statische Dokumentenmenge
- **Gegeben:** Bedürfnis nach Information (keyword-based query)
- **Aufgabe:** finde alle und nur die Dokumente, die für die query relevant sind

Typische IR Systeme:

- Suche in einer Menge von Abstracts
- Suche in Zeitungsartikeln
- Bibliotheksssuche
- Websuche

Themen in IR

- Wie formuliere ich eine query? (Query type)
- Wie sind die Dokumente repräsentiert? (indexing)
- Wie findet das System die passenden Dokumente? (retrieval model)
- Effizienz?
- Wie werden die Dokumente präsentiert? (unsortierte Liste, geordnete Liste, clusters)
- Wie evaluiere ich das System? (Evaluation)

Indizierung

Finde Terme, die ein Dokument gut beschreiben

- Manuell:
  - Indizierung durch Menschen mit fixem Vokabular
  - Arbeits- und Trainingsintensiv
- Automatisch:
  - Term manipulation
    (einige Worte als selber Term)
  - Termgewichtung
  - Indexterme alle aus dem Text

Manuelle Indizierung

- Großes Vokabular (mehr als 1K)
  - ACM – Themen in Computer Science
  - Library of Congress Subject Headings
- Probleme:
  - Indizierer brauchen Training
  - Dokumente dynamisch $\rightarrow$ Vokabular muss sich ändern
- Vorteile:
  - Hohe Akkuratheit bei Suche
  - Gut für geschlossene Sammlungen

Beispiel

Medical Subject Headings (MeSH)	
...	
Eye Diseases	C11
Asthenopia	C11.93
Conjunctival Diseases	C11.187
Conjunctival Neoplasms	C11.187.169
Conjunctivitis	C11.187.183
Conjunctivitis, Allergic	C11.187.183.200
Conjunctivitis, Bacterial	C11.187.183.220
Conjunctivitis, Inclusion	C11.187.183.220.250
Ophthalmia Neonatorum	C11.187.183.220.538
Trachoma	C11.187.183.220.889
Conjunctivitis, Viral	C11.187.183.240
Conjunctivitis, Acute Hemorrhagic	C11.187.183.240.216
Keratoconjunctivitis	C11.187.183.394

Beispiel

Computing Classification System (1998)
B Hardware
B.3 Memory structures
B.3.0 General
B.3.1 Semiconductor Memories (NEW) (was B.7.1)
Dynamic memory (DRAM) (NEW)
Read-only memory (ROM) (NEW)
Static memory (SRAM) (NEW)
B.3.2 Design Styles (was D.4.2)
Associative memories
Cache memories
Interleaved memories
Mass storage (e.g., magnetic, optical, RAID)
Primary memory
Sequential-access memory

Automatische Indizierung

- Keine vorher festgelegte Indizierungsterme
- Benutze Worte im Dokument
- Implementierung der Indizes: inverted files

Inverted files

Document	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

Inverted files

Document	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

Number	Text	Documents
1	cold	1, 4
2	days	3, 6
3	hot	1, 4
4	in	2, 5
5	it	4, 5
6	like	4, 5
7	nine	3, 6
8	old	3, 6
9	pease	1, 2
10	porridge	1, 2
11	pot	2, 5
12	some	4, 5
13	the	2, 5

Inverted files: Mit Position

Document	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

$\Longrightarrow$

Number	Text	(Document; Word)
1	cold	(1; 6), (4; 8)
2	days	(3; 2), (6; 2)
3	hot	(1; 3), (4; 4)
4	in	(2; 3), (5; 4)
5	it	(4; 3, 7), (5; 3)
6	like	(4; 2, 6), (5; 2)
7	nine	(3; 1), (6; 1)
8	old	(3; 3), (6; 3)
9	pease	(1; 1, 4), (2; 1)
10	porridge	(1; 2, 5), (2; 2)
11	pot	(2; 5), (5; 6)
12	some	(4; 1, 5), (5; 1)
13	the	(2; 4), (5; 5)

Indizierung: Diskussion

- Vorteile und Nachteile davon, so viele Seiten wie möglich zu indizieren ?
- Dynamisch generierte Seiten?
- Dynamische Indizierung

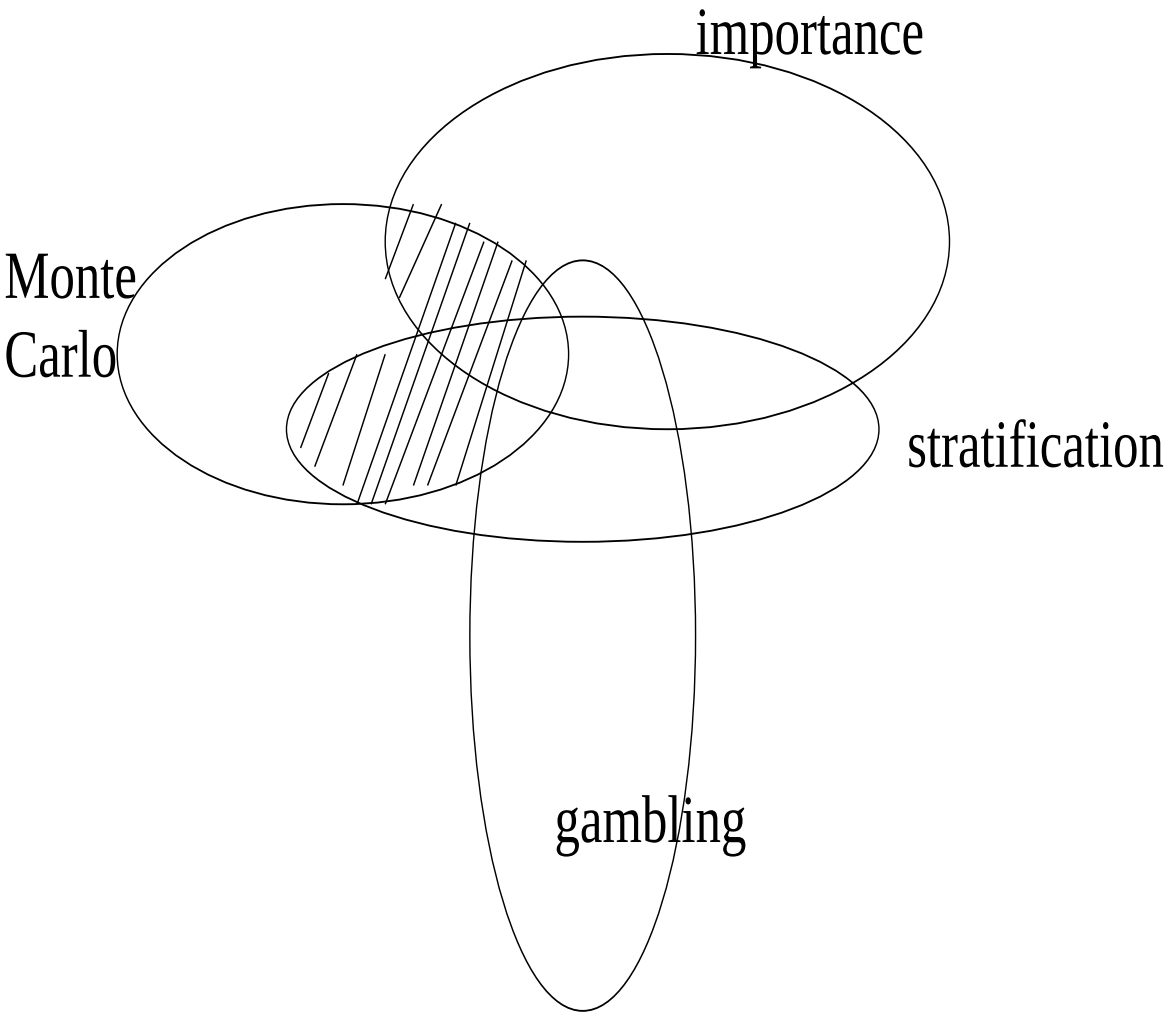
Siehe auch: Henzinger et al.: Search Technologies for the Internet in *Science 317*, 2007 sowie für Index construction und Index compression auch Manning et al (2008) *Introduction to Information Retrieval*.

Information Retrieval: Methoden

- **Boolsche Suche:**
  - binäre Entscheidung
  - Präsenz des Terms notwendig und ausreichend für Match
  - Boolsche Mengenoperatoren (AND, OR)
- **ranked algorithms:**
  - Frequenz der Dokumentterme
  - nicht alle Suchterme unbedingt im Dokument
  - Beispiele:
    - **Vektorraummodell**
      (SMART, Salton et al, 1971)
    - Probabilistisches Modell
      (OKAPI, Robertson/Spärck Jones, 1976)
    - Websuchmaschinen

Das Boolsche Modell

Monte Carlo AND (importance OR stratification) BUT gambling



Das Boolsche Modell

Monte Carlo AND (importance OR stratification) BUT gambling

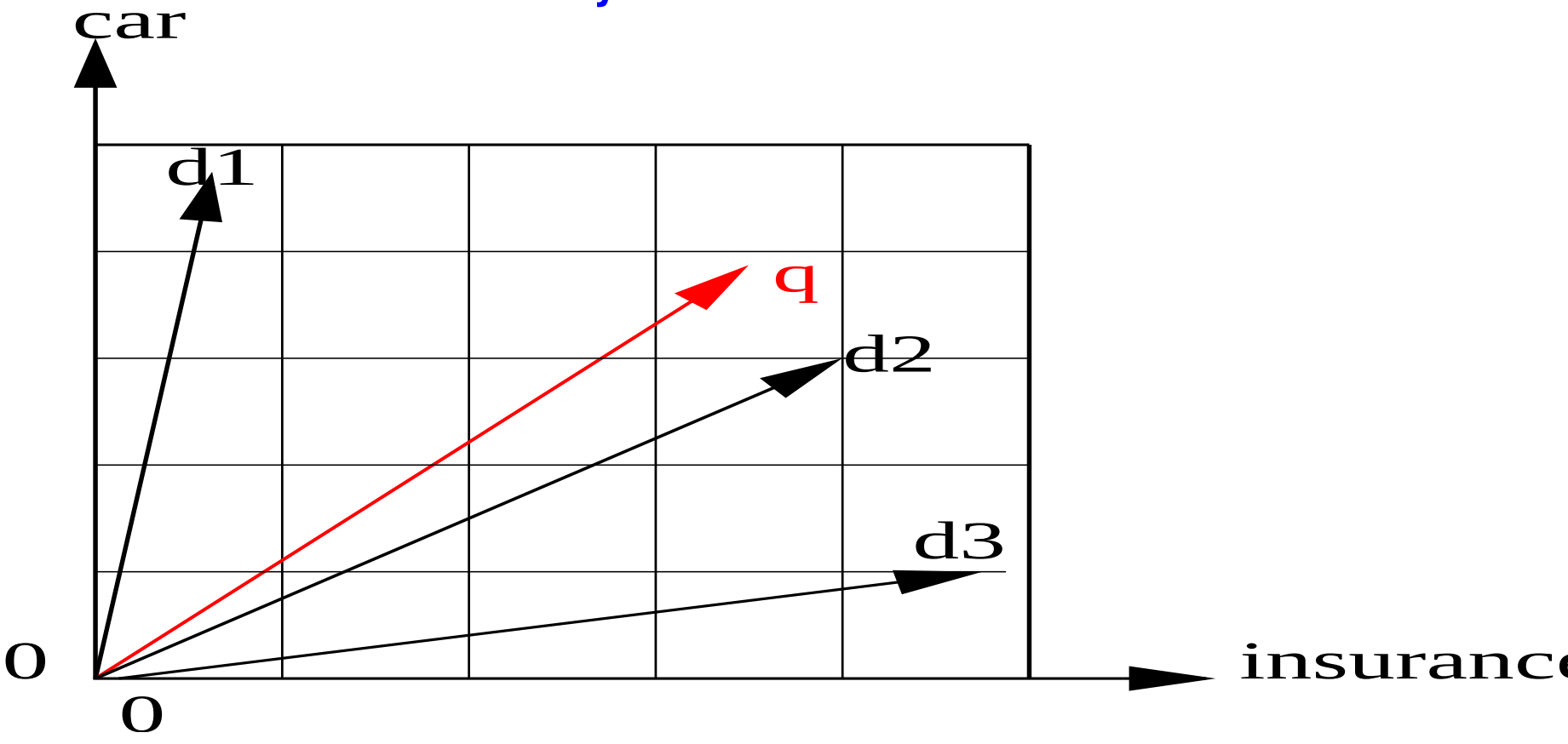
- Mengentheoretische Interpretation von AND OR BUT
- Typisch für Bibliothekssuche
- **Problem 1:** Expertenwissen für gute Suche
- **Problem 2:** Binäre Relevanzdefinition $\rightarrow$ ungeordnete Resultatlisten

The Vector Space model

- Dokumente Vektoren in hochdimensionalem Raum
- Queries ebenfalls
- Document–query Ähnlichkeit bestimmt Relevanz (ranking)

The Vector Space model

Query: car insurance



Vektorraummodell: Wort-Dokument-matrix

	Term_1	Term_2	Term_3	...	Term_n
Doc_1	14	6	1	...	0
Doc_2	0	1	3	...	1
Doc_3	0	1	0	...	2
...	...	...	...	...	...
Doc_N	4	7	0	...	5
Q	0	1	0	...	1

- Vektorähnlichkeitsmaß
- Was ist ein Term?
- Sind Terme gewichtet?

Ähnlichkeit

$$\vec{x} = (x_1, x_2, \ldots, x_n)$$

Der Kosinus **cosine** des Winkels

$$cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

Cosinus das meist gewählte Ähnlichkeitsmaß in IR!

Was ist ein Term?

- Alle Worte?

Was ist ein Term?

- Alle Worte?
- Normalisierung? (**Turkey** – **turkey**)

Was ist ein Term?

- Alle Worte?
- Normalisierung? (**Turkey** – **turkey**)
- Phrasen (**cheque book** – **cheque** und **book**)

Was ist ein Term?

- Alle Worte?
- Normalisierung? (**Turkey** – **turkey**)
- Phrasen (**cheque book** – **cheque** und **book**)
- Stoplist

Was ist ein Term?

- Alle Worte?
- Normalisierung? (**Turkey** – **turkey**)
- Phrasen (**cheque book** – **cheque** und **book**)
- Stoplist

a	always	both
about	am	being
above	among	co
across	amongst	could
after	are	done

Was ist ein Term?

- Alle Worte?
- Normalisierung? (**Turkey** – **turkey**)
- Phrasen (**cheque book** – **cheque** und **book**)
- Stoplist
- Stemmer

Was ist ein Term?

- Alle Worte?
- Normalisierung? (**Turkey** – **turkey**)
- Phrasen (**cheque book** – **cheque** und **book**)
- Stoplist
- Stemmer

CONNECT	WORRY	GALL
CONNECT**ED**	WORR**IED**	GALL**ING**
CONNECT**ING**	WORR**IES**	GALL**ED**
CONNECT**ION**	WORRY**ING**	GALL**EY**
CONNECT**IONS**	WORRY**INGLY**	GALL**ERY**

Termgewichtung

- Frequenz?

term freq	$tf_{w,d}$	Häufigkeit von w_i in d_j
document freq	df_w	Anzahl der Dokumente mit w
document collection	$\|D\|$	Anzahl der Dokumente

Wort	$tf_{i,j}$	df
insurance	20	3997
try	5	9760

Termgewichtung: TF/IDF

Terme, die oft in einem Dokument vorkommen, beschreiben das Dokument besser:

$$tf_{w,d} = freq_{w,d}$$

Terme, die in den andere Dokumenten selten sind, beschreiben das Dokument besser:

$$idf_{w,D} = log \frac{|D|}{df_w}$$

Kombiniere *tf* mit *idf*:

$$tfidf_{w,d,D} = tf_{w,d} \cdot idf_{w,D}$$

Beispiel: TF/IDF

Term	tf	df_w	$\lvert D \rvert$	TF/IDF
the	312	28,799	30,000	
in	179	26,452	30,000	
general	136	179	30,000	
fact	131	231	30,000	
explosives	63	98	30,000	
nations	45	142	30,000	
haven	37	227	30,000	

Beispiel: TF/IDF

Term	tf	df_w	$\lvert D \rvert$	TF/IDF
the	**312**	**28,799**	**30,000**	
in	179	26,452	30,000	
general	136	179	30,000	
fact	131	231	30,000	
explosives	63	98	30,000	
nations	45	142	30,000	
haven	37	227	30,000	

$$IDF(the) = \log(\frac{30,000}{28,799}) = 0.0178$$

$$TF/IDF(the) = 312 \cdot 0.0178 = 5.55$$

Beispiel: TF/IDF

Term	tf	df_w	$\lvert D \rvert$	TF/IDF
the	312	28,799	30,000	**5.55**
in	179	26,452	30,000	**9.78**
general	136	179	30,000	**302.50**
fact	131	231	30,000	**276.87**
explosives	63	98	30,000	**156.61**
nations	45	142	30,000	**104.62**
haven	37	227	30,000	**78.48**

Beispiel: VSM, TF/IDF, Kosinus

	Q	D_{7655}	D_{454}
hunter	19.2	56.4	112.2
gatherer	34.5	122.4	0
Scandinavia	13.9	0	30.9
30,000	0	457.2	0
years	0	12.4	0
BC	0	200.2	0
prehistoric	0	45.3	0
deer	0	0	23.6
rifle	0	0	452.2
Mesolithic	0	344.2	0

$cos(Q, D_{7655}) = .20$

$cos(Q, D_{454}) = .13$

Evaluation

Gebraucht wird:

- Dokumentenmenge
- Test suite von Informationsbedürfnissen, die als “queries” ausgedrückt werden können (mind 50)
- (binäre) Relevanzurteile für Goldstandard

Datensätze

- **Cranfield:**
  - 1398 Abstracts von Aerodynamikartikeln
  - 225 queries
  - Relevanzurteile für alle query-Dokumentpaare
- TREC jährliche Evaluationen
  - über 500,000 Zeitungstexte
  - 150 queries
  - Relevanzurteile via **pooling**

Pooling

Relevanzurteile über Teilmenge der Texte pro query. Teilmenge besteht aus der Sammlung der Top *n* Texten, die Systeme zurückgegeben haben.

Evaluation: Maße

	Relevant	**Non-relevant**	**Total**
Retrieved	A	B	A+B
Not retrieved	C	D	C+D
Total	A+C	B+D	A+B+C+D

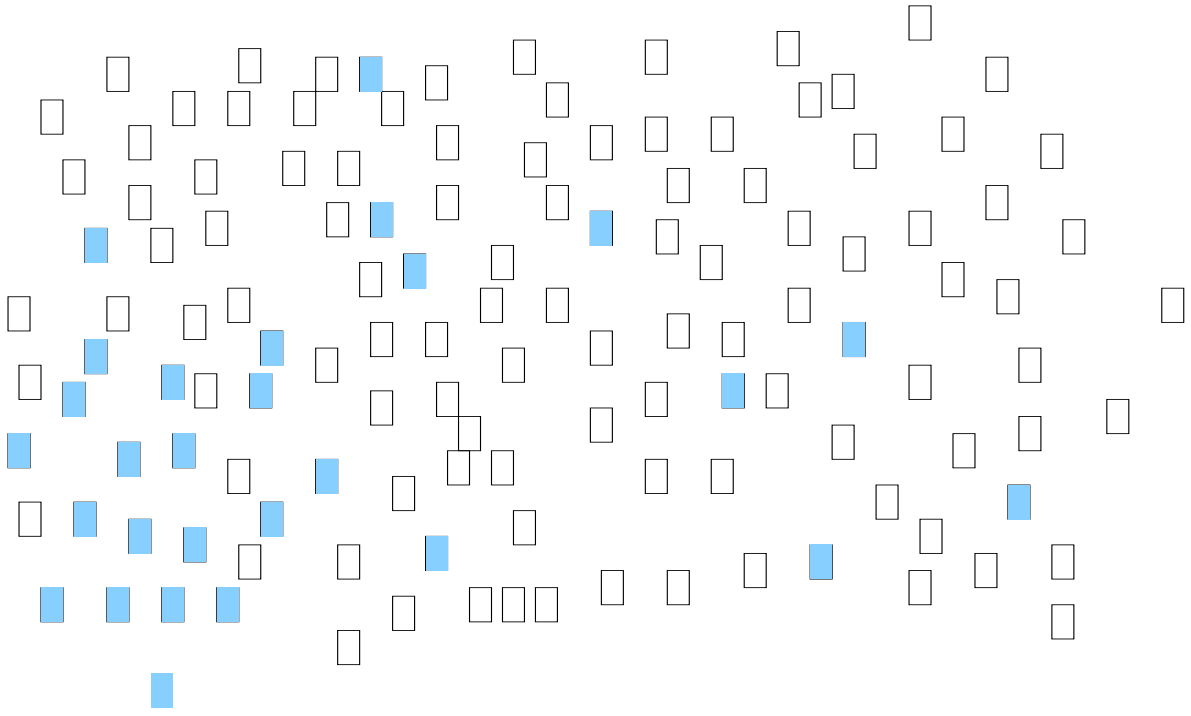
Recall $\frac{A}{A+C}$

Precision $\frac{A}{A+B}$

Accuracy $\frac{A+D}{A+B+C+D}$

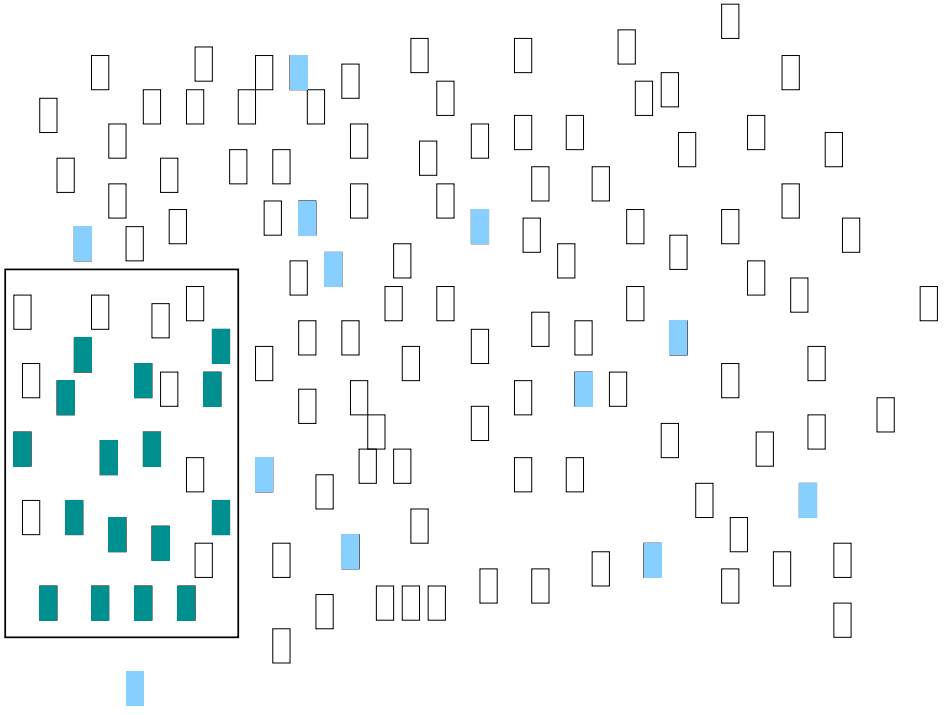
Accuracy normalerweise nicht für IR benutzt. Warum?

Recall und Precision



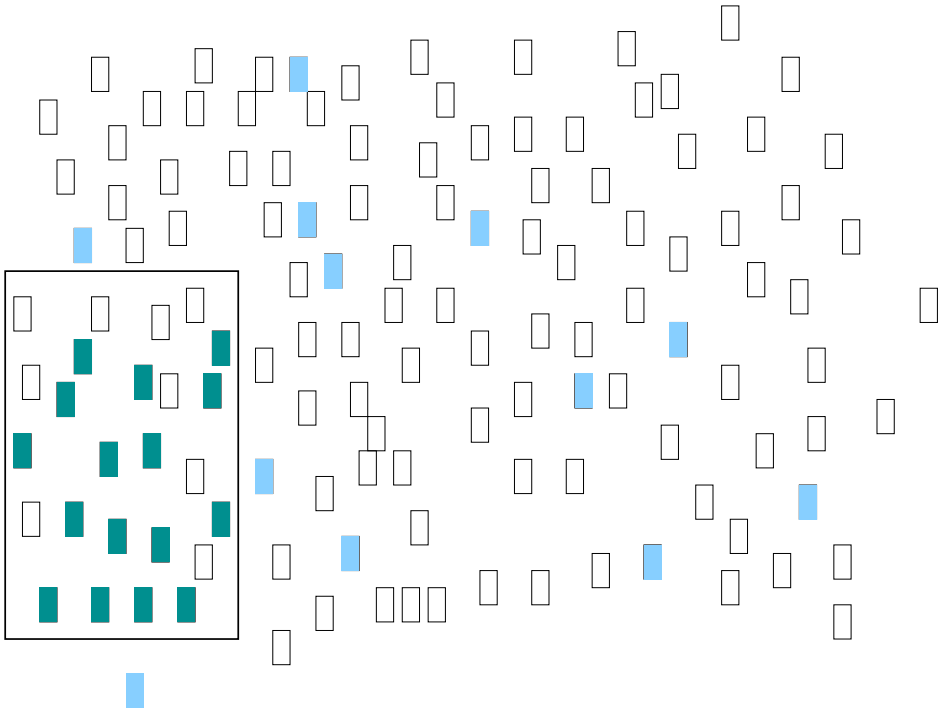
- Alle Dokumente: A+B+C+D = 130
- Relevante Dokumente: A+C = 28

Recall und Precision: System



- System 1 findet 25 Dokumente: $(A+B)_1 = 25$
- Relevante und gefundene Dokumente: $A_1 = 16$
- Relevante Dokumente: $A+C = 28$

Recall und Precision: System



$$R_1 = \frac{A_1}{A+C} = \frac{16}{28} = .57$$
$$P_1 = \frac{A_1}{(A+B)_1} = \frac{16}{25} = .64$$

- System 1 findet 25 Dokumente: $(A+B)_1 = 25$
- Relevante und gefundene Dokumente: $A_1 = 16$
- Relevante Dokumente: $A+C = 28$

F-measure

Maß, dass Precision und Recall verbindet Rijsbergen (1970).

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

- Hohes α: Recall wichtiger
- Niedriges α: Precision wichtiger
- Meist α=0.5
- Gewichteter harmonischer Mittelwert P and R

$$F_{0.5} = \frac{2PR}{P+R}$$

Eigenschaften Precision, Recall, F-measure

- Mengenbasiert
- Evaluieren Ranking nicht
- Websuche: Precision wichtiger
- Manche andere Suchprobleme (z.B. Suche bei Rechtsproblemen): Recall sehr wichtig

Rankingevaluation

Evaluation	Ranking 1		Ranking 2		Ranking 3	
	d1:	√	d10:	×	d6:	×
	d2:	√	d9:	×	d1:	√
	d3:	√	d8:	×	d2:	√
	d4:	√	d7:	×	d10:	×
	d5:	√	d6:	×	d9:	×
	d6:	×	d1:	√	d3:	√
	d7:	×	d2:	√	d5:	√
	d8:	×	d3:	√	d4:	√
	d9:	×	d4:	√	d7:	×
	d10:	×	d5:	√	d8:	×

Welches System ist das Beste?

Precision an einem Recall-Level

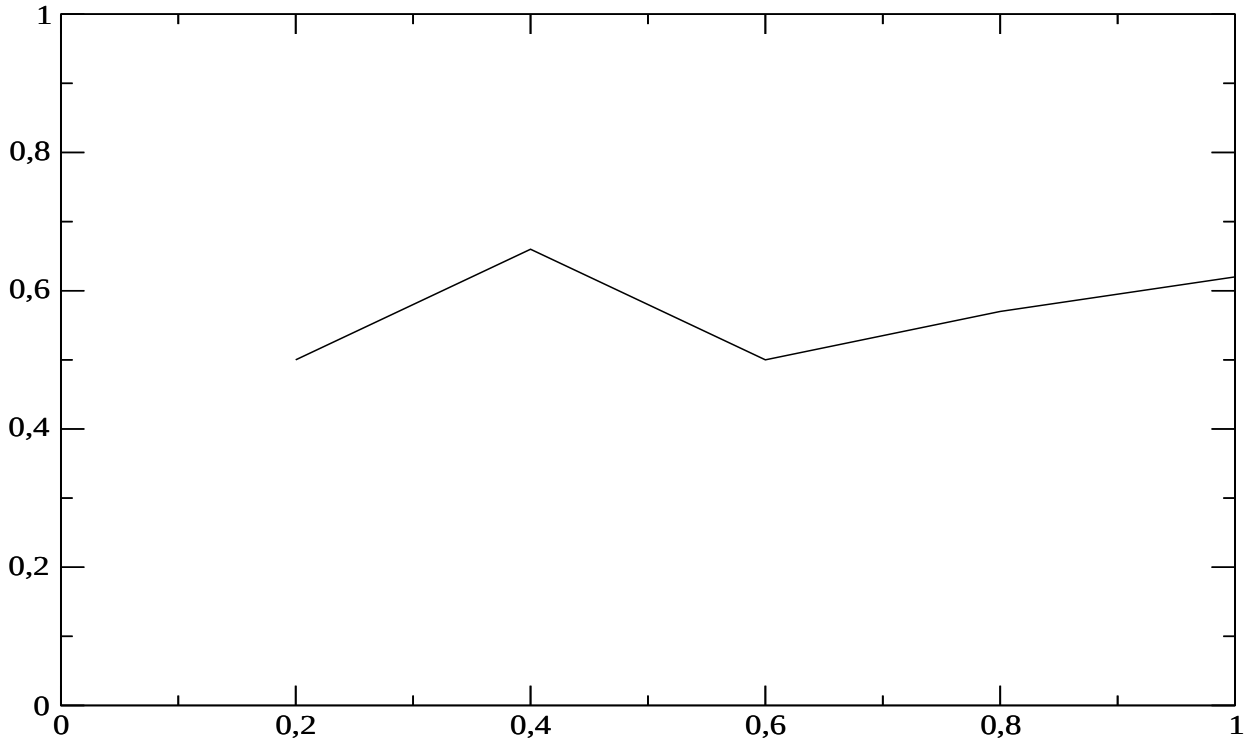
Annahme: genau 5 Dokumente waren relevant insgesamt.

Ranking 3		Precision	Recall	RLevel
d6:	×	0	0	
d1:	√	1/2	1/5	20%
d2:	√	2/3	2/5	40%
d10:	×	0	0	
d9:	×	0	0	
d3:	√	3/6	3/5	60%
d5:	√	4/7	4/5	80%
d4:	√	5/8	5/5	100%
d7:	×	0	0	
d8:	×	0	0	

Precision at Recall level 1: $Prec(R_1) = 0.5$

Precision-Recall Kurven

Zeigen alle “Precision an einem Recall-Level”:



Average Precision

Bildet den Durchschnitt aller “Precision at Recall Level” Werte

Wenn die query m relevante Dokumente hat, dann

$$AP(q) = \frac{1}{m} \sum_{1 \leq i \leq m} Prec(R_i)$$

Average Precision

Evaluation	Ranking 1		Ranking 2		Ranking 3	
	d1:	$\surd$ (1/1)	d10:	$\times$	d6:	$\times$
	d2:	$\surd$ (2/2)	d9:	$\times$	d1:	$\surd$ (1/2)
	d3:	$\surd$ (3/3)	d8:	$\times$	d2:	$\surd$ (2/3)
	d4:	$\surd$ (4/4)	d7:	$\times$	d10:	$\times$
	d5:	$\surd$ (5/5)	d6:	$\times$	d9:	$\times$
	d6:	$\times$	d1:	$\surd$ (1/6)	d3:	$\surd$ (3/6)
	d7:	$\times$	d2:	$\surd$ (2/7)	d5:	$\surd$ (4/7)
	d8:	$\times$	d3:	$\surd$ (3/8)	d4:	$\surd$ (5/8)
	d9:	$\times$	d4:	$\surd$ (4/9)	d7:	$\times$
	d10:	$\times$	d5:	$\surd$ (5/10)	d8:	$\times$
avg. prec	1.0		0.3544		0.5726	

Mean Average Precision

Gegeben eine Menge von queries Q, wobei eine query q_j m_j relevante Dokumente hat.

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1\ldots|Q|} \left(\frac{1}{m_j} \sum_{1 \leq i \leq m_j} Prec(R_i^j)\right)$$

- Meistbenutztes Evaluationsmaß
- Keine festgelegten Recall-Level, alle werden benutzt
- Man muss für AP jeweils alle relevanten Dokumente kennen!
- Jede query gleich wichtig
- Da AP per query stark variieren kann, müssen viele benutzt werden

Precision at a cutoff

Evaluation	Ranking 1		Ranking 2		Ranking 3	
	d1:	√	d10:	×	d6:	×
	d2:	√	d9:	×	d1:	√
	d3:	√	d8:	×	d2:	√
	d4:	√	d7:	×	d10:	×
	d5:	√	d6:	×	d9:	×
	d6:	×	d1:	√	d3:	√
	d7:	×	d2:	√	d5:	√
	d8:	×	d3:	√	d4:	√
	d9:	×	d4:	√	d7:	×
	d10:	×	d5:	√	d8:	×
Precision at 5	1.0		0.0		0.4	

Precision at a cutoff

Evaluation	Ranking 1		Ranking 2		Ranking 3	
	d1:	√	d10:	×	d6:	×
	d2:	√	d9:	×	d1:	√
	d3:	√	d8:	×	d2:	√
	d4:	√	d7:	×	d10:	×
	d5:	√	d6:	×	d9:	×
	d6:	×	d1:	√	d3:	√
	d7:	×	d2:	√	d5:	√
	d8:	×	d3:	√	d4:	√
	d9:	×	d4:	√	d7:	×
	d10:	×	d5:	√	d8:	×
Precision at 10	0.5		0.5		0.5	

Precision at cutoff k

Eigenschaften:

- Benutzt nur Top k zurückgegebene Dokumente
- Gibt es viele relevante Dokumente, dann kann dies Performanz überschätzen
- Braucht nicht alle relevanten Dokumente zu kennen
- Wichtig für Websuche: hohe Präzision auf erster Seite erwünscht

Kritik an Relevanzerhebung

- Relevanzerhebung eines Dokuments unabhängig von anderen Dokumenten $\rightarrow$ search diversity?
- Binär

Ist Relevanz wirklich alles was User interessiert?

- Geschwindigkeit von Indizierung
- Geschwindigkeit der Suche
- Snippets
- Web page design

Dies alles wird gemessen in

- User studies
- Wie oft kommt ein User zurück?
- Wieviele Benutzer werden Käufer?
- **A/B testing**: Ändere nur einen Parameter für eine kleine Prozentzahl der Nutzer, und sehe, wie sich das Clickverhalten ändert (clickstream mining). Auch für Relevanz

Zwischenfazit

- Information retrieval
- IR modelle
  - Boolsche Modelle
  - Vektorraummodelle
- Termmanipulation und TF/IDF
- Evaluation

Referenzen:

- Manning et al (2008) *Introduction to IR*

Noch ungenügend

Für heutige Websuche ist dieses simple Vektorraummodell ungenügend, da die Menge der Dokumente zu groß ist, das Ranking nicht nur von den Dokumenten abhängt etc. Da das Web ein Graph ist, habe wir auch weitere Möglichkeiten.

- Popularität von Dokumenten für das Ranking: PageRank nützt den Webgraphen aus
- Spam
- Filters
- Userinteressen: long-term, short-term
- Personalisierung

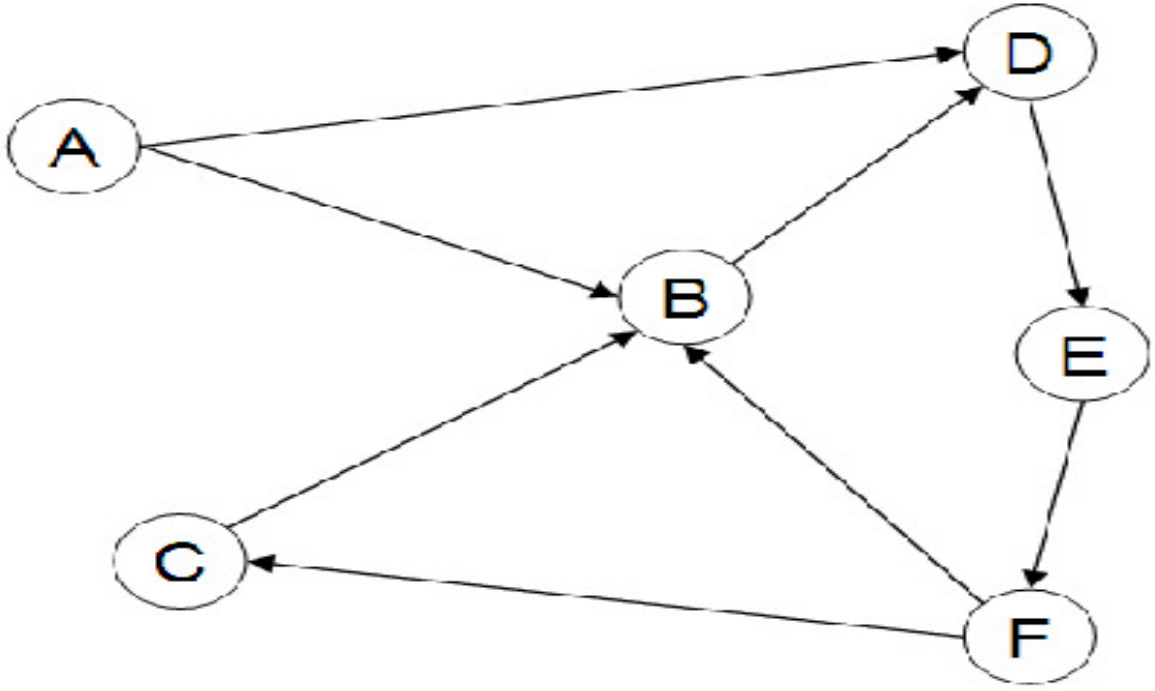
PageRank

- Ein statisches, query-unabhängiges Qualitätsmerkmal einer Webseite
- Beruht auf dem Webgraphen, nicht dem Webtext
- Approximiert Popularität von Seiten

Graphen

Graph

Besteht aus Knoten $\{V\}$ und Kanten $\{E\} \subseteq V \times V$. Kanten können gerichtet sein.

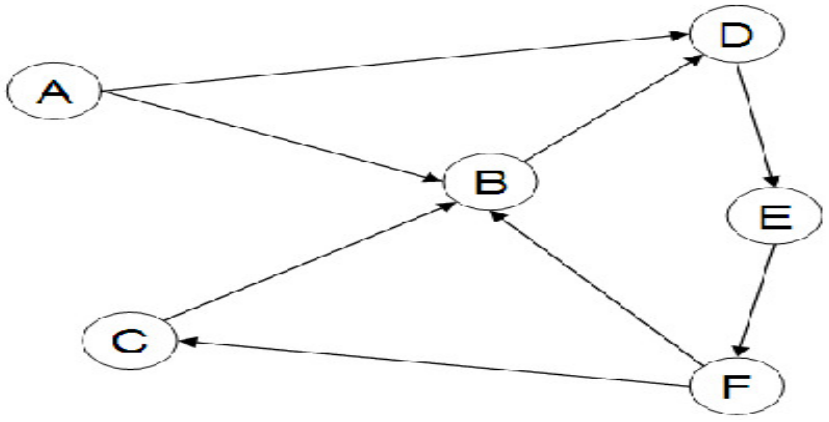


Web: Knoten = Seiten, Kanten: links zwischen Seiten

Grafik aus Manning et al. Introduction to IR

Eigenschaften

Grafik aus Manning et al. Introduction to IR



- Out-degree von A: 2
- In-degree von A: 0
- Webgraphen: meist nicht stark verbunden (stark verbunden = man kann von jedem Knoten zu jedem anderen gelangen)
- Web hat Zipfsche Verteilung, d.h. Anzahl der Seiten mit in-Links i proprotional zu $\frac{1}{i^{\alpha}}$

Random Walk

- Surfer startet irgendwo an einem der N Knoten v_0
- Folgt mit Wahrscheinlichkeit $1-\alpha$ zufällig einem der k outlinks von v_0 zum nächsten Knoten mit gleicher Wahrscheinlichkeit. Also Wahrscheinlichkeit für jeden Outlink $(1-\alpha)\frac{1}{k}$.
- Mit Wahrscheinlichkeit α, beamt er stattdessen zu irgendeinem Knoten. Wahrscheinlichkeit jedes Knotens $\alpha\frac{1}{N}$
- Gibt es keine Outlinks, kann er nur “beamen” . Teleportation mit Wahrscheinlichkeit $1/N$ für jeden Knoten
- Mache so weiter am nächsten besuchten Knoten

Beispiel: Teil 1

Verbindungen unter drei Knoten

- $A \rightarrow B$
- $C \rightarrow B$
- $B \rightarrow A$
- $B \rightarrow C$

Als **adjacency matrix**, wobei jede Verbindung mit 1 gekennzeichnet ist:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Übergangswahrscheinlichkeiten

Verwandlung in Matrix von **Übergangswahrscheinlichkeiten** eines Markov-Prozesses.

- Wenn eine Zeile r_i keine Einsen hat (keine outlinks), dann addiere zu jedem Eintrag $1/N$. Fertig!
- **Sonst:** Eintrag = $(1-\alpha)\frac{r_{i,j}}{\sum_j r_{i,j}} + \alpha \cdot \frac{1}{N}$
  - Teile jeden Eintrag durch Anzahl der Einsen in Zeile (random walk)
  - Dann: Multipliziere alle Einträge mit $1-\alpha$ (walk mit Wahrscheinlichkeit $1-\alpha$)
  - Addiere α/N zu allen Einträgen (teleport)

Beispiel: Teil 2

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

wird (mit α =0.5) zu T

$$\begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}$$

Eigenschaften

- **Stochastische Matrix:** Matrix mit nicht-negativen Einträgen, in der sich alle Zeileneinträge zu 1 addieren
- **irreduzibel**: Sequenz von Wahrscheinlichkeiten grösser Null von jdem Zustand zu jedem anderen
- **aperiodisch**: man wird nicht in Zyklen gefangen

Man kann zeigen:

Wenn der random walker lange läuft, dann konvergiert die Wahrscheinlichkeit der Besuchshäufigkeiten der Web pages unabhängig vom Anfang des walks. Dies ergibt den Page Rank einer Seite. D.H man kann mit jedem stochastischem Vektor beginnen

Beispiel: Teil 3

Nehme an Walker startet in A: $v_0 = (1,0,0)$

Schritt 1 Wahrscheinlichkeiten

$$v_0 \cdot T = (1,0,0) \cdot \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix} = (1/6, 2/3, 1/6)$$

Beispiel: Teil 3

Weitere Schritte:

v_0	1	0	0
v_1	1/6	2/3	1/6
v_2	1/3	1/3	1/3
v_3	1/4	1/2	1/4
v_4	7/24	5/12	7/24
...			
v	5/18	4/9	5/18

v ist der page rank oder steady-state vector und gibt die mittlere Besuchshäufigkeit bei unendlicher Laufzeit an. Es gilt $v \cdot T = v$

PageRank

- Man bekommt höheren Score, wenn man viele In-Links hat
- die auch wieder von wichtigen Seiten mit vielen In-Links stammen
- α oft 0.1
- Wird kombiniert mit Kosinusähnlichkeit als query-unabhängiges Qualitätsmerkmale
- Andere Qualitätsmerkmale: Dokumentlänge, Dokumentalter ...
- Random Walk nicht nur für Websuche
- Manning et al: Introduction to IR. Kapitel 20.1