

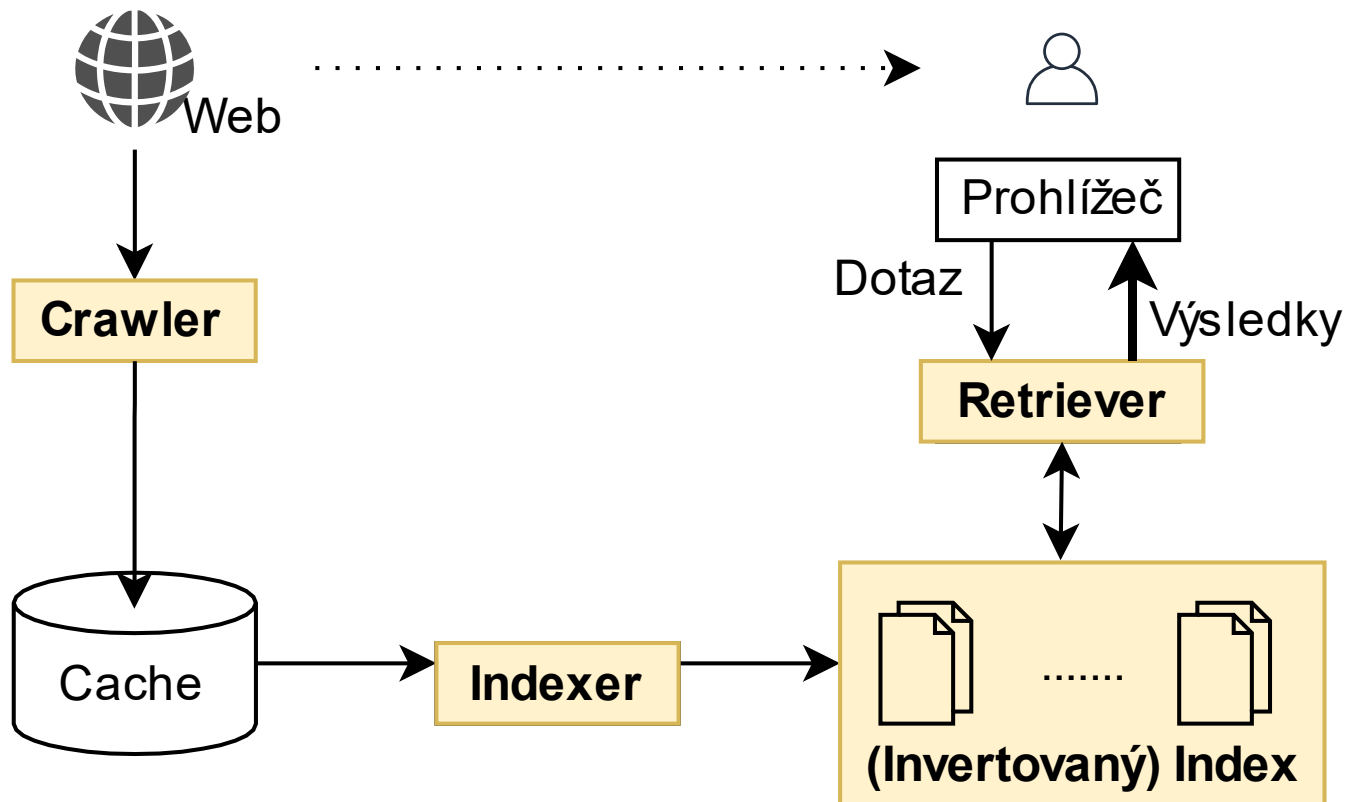
Hodnocení a vyhledávání na webu

František Kynych
11. 11. 2021 | MVD



Část I.: Úvod do problematiky, základní části systému

Co je to vyhledávač?



Výzvy a možnosti vyhledávání

- Škálovatelnost
 - Zvládnutí velikosti webu
 - Zaručit úplnost pokrytí webu
 - > **Paralelní indexování a vyhledávání (MapReduce)**
- Stránky s nízkou kvalitou informací a spam
 - > **Detekce spamu a spolehlivé (robustní) hodnocení**
- Dynamika webu
 - Nepřetržitá tvorba nových stránek
 - Časté aktualizace některých stránek
- Možnosti
 - Mnoho dalších heuristik (např. linky) může být využito k vylepšení vyhledávání
 - > **Analýza linků a využití více příznaků k hodnocení**

Crawler / Spider / Robot

- Bot, který prochází web za účelem vytvoření web indexu
- Snadné vytvoření základního bota
 - Začneme s kolekcí několika stránek v prioritní frontě
 - Načteme stránku z webu
 - Z načtených stránek parsujeme linky a přidáváme je do fronty
 - Procházíme celou frontu
- Reálný bot je více komplikovaný
 - Robustnost (chyba serveru, pasti, ...)
 - Pravidla pro bota (ohled na vytížení serverů, některé stránky si nepřejí být indexovány, ...)
 - Práce s různými typy souborů (dokumenty, obrázky, skripty, ...)
 - Odhalení redundantních stránek (identické / duplicitní)
 - Odhalení skrytých URL (zkrácení dlouhé URL)

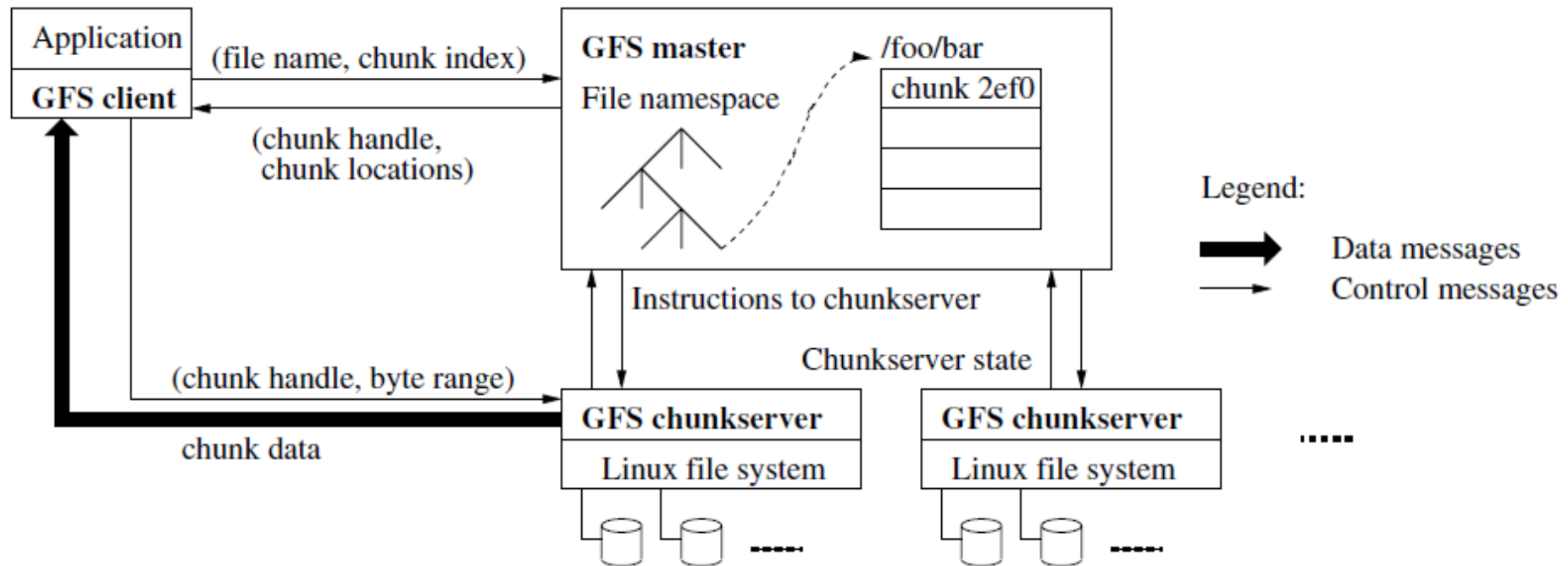
Základní strategie prohledávání

- Prohledávání do šířky (většinou)
 - Chceme vyvážit zatížení serveru
- Paralelní prohledávání
- Variace: zaměřené prohledávání
 - Zaměřujeme se pouze na jedno téma (např. stránky o sportu)
 - Můžeme zadat dotaz do již existujícího vyhledávače a procházet výsledky
- Problém s hledáním nových stránek
- Inkrementální / opakované prohledávání
 - Chceme minimalizovat použití výpočetních zdrojů
 - Bot se může učit ze zkušeností (procházet denně / měsíčně)
 - Může zaměřovat
 - Často aktualizované stránky
 - Často procházené stránky uživateli

Indexování

- Standardní techniky z přednášky o vyhledávání jsou základem, ale nejsou dostačující
 - Škálovatelnost
 - Efektivnost
- Důležité části
 - Google File System (GFS)
 - Distribuovaný souborový systém
 - MapReduce
 - Softwarový model pro paralelní zpracování dat
 - Hadoop
 - Obsahuje open source implementaci MapReduce

Google File System (GFS)

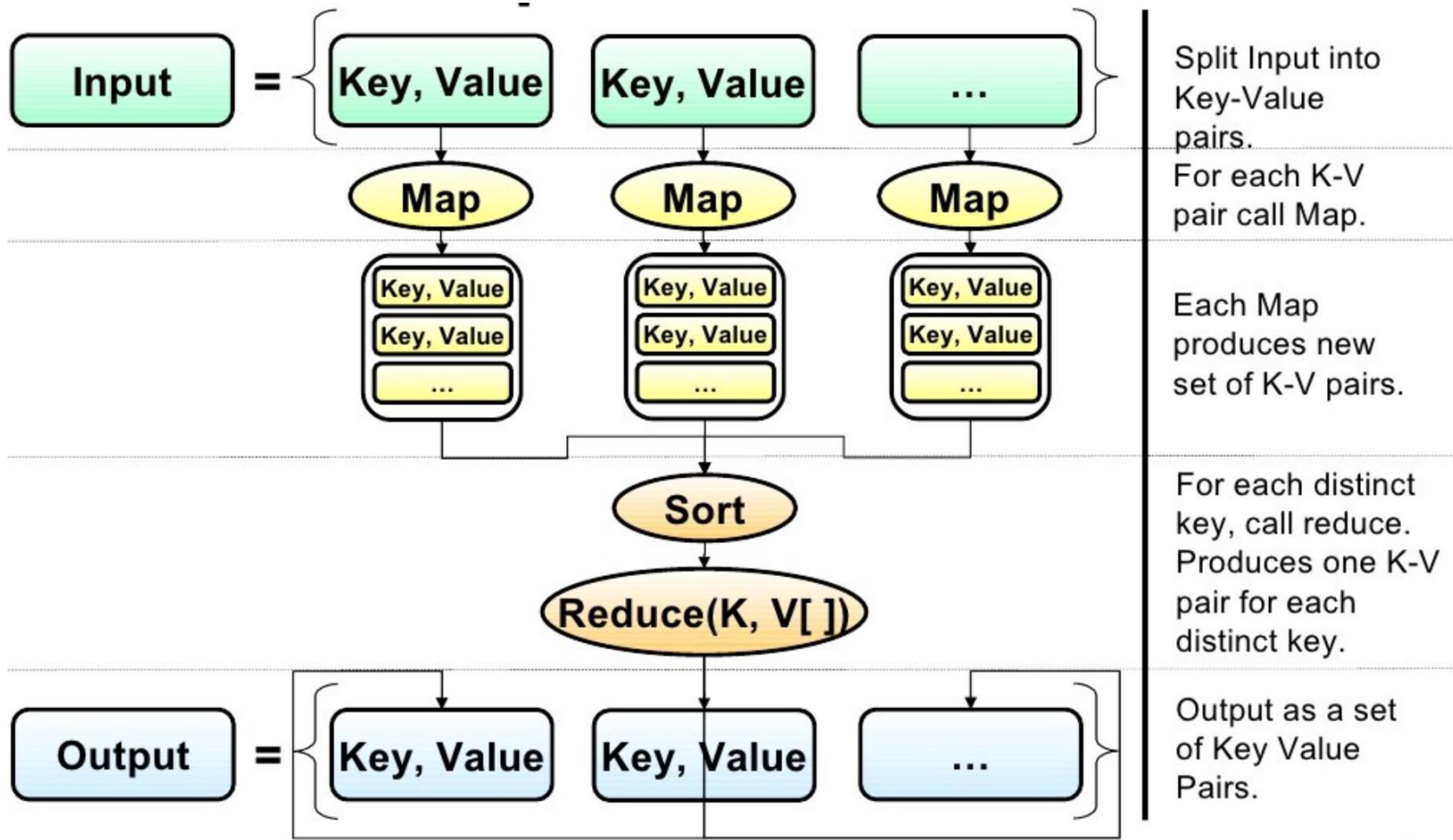


<https://static.googleusercontent.com/media/research.google.com/cs/archive/gfs-sosp2003.pdf>

MapReduce

- Model pro paralelní zpracování velkého množství dat
 - Minimalizuje potřebné úsilí programátora pro jednoduché úlohy paralelního zpracování dat
- U indexování zpracovává uložené webové stránky z GFS
- Výhody
 - Skrývá před programátorem většinu low-level detailů (sít', úložiště)
 - Odolnost proti chybám
 - Automatické vyvažování zátěže

MapReduce výpočetní pipeline

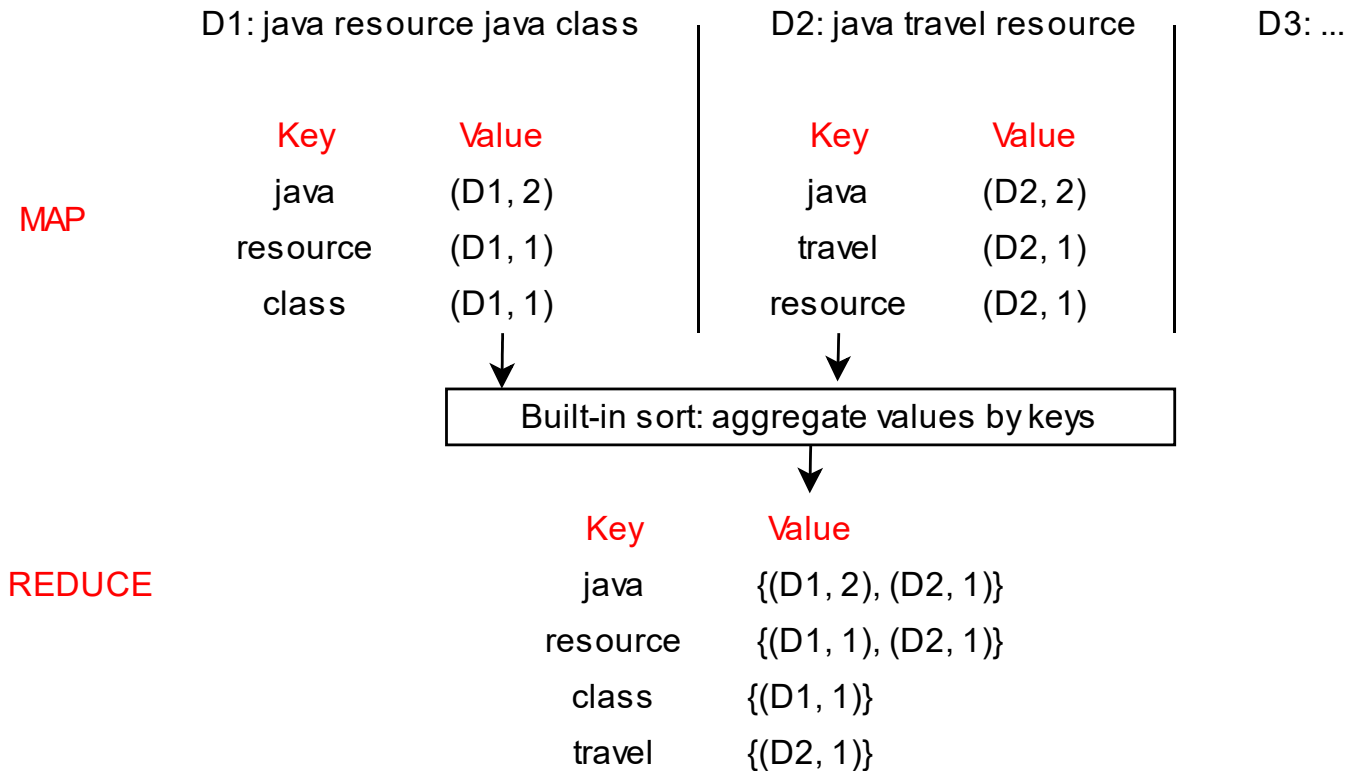


<https://www.slideshare.net/gothicane/behm-shah-pagerank>



MapReduce příklad

- Získání dat pro invertovaný index pomocí MapReduce



Apache Hadoop

- Framework obsahující open source komponenty
 - Hadoop Distributed File System (HDFS)
 - Distribuovaný souborový systém pro rychlý přístup k velkým datům
 - Hadoop YARN
 - Framework pro plánování jobů a správu zdrojů clusterů
 - **Hadoop MapReduce**
 - Systém pro paralelní zpracování velkého množství dat

Část II.: Algoritmy pro hodnocení stránek

Hodnocení stránek

- Standardní modely z předchozí přednášky fungují, ale pro tuto úlohu nestačí
 - Uživatel má při vyhledávání na webu jiné potřeby
 - Dokumenty obsahují více informací (layout, linky, ...)
 - Výrazně se liší kvalita informací na stránce
- Hlavní rozšíření
 - Využití linků k vylepšení hodnocení
 - Využití návštěvnosti pro přímou zpětnou vazbu
 - Využití machine learning algoritmů pro kombinaci různých příznaků

Využití linků mezi stránkami

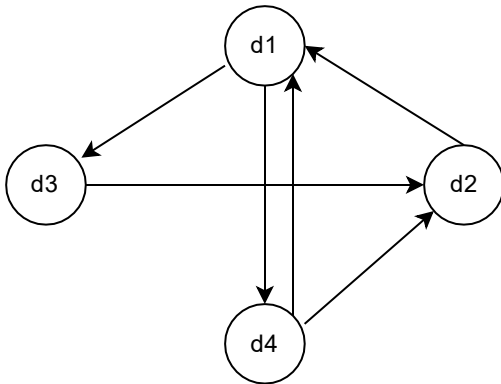
- Odkazy ze stránek
 - Většinou má link nějaký text -> víte kam přecházíte
 - Tento text lze využít
 - Odkaz „[the biggest online book store](#)“ převede uživatele např. na Amazon
 - Jakmile uživatel bude vyhledávat podobný text, tak těchto linků můžeme využít
- Hub stránky
 - Stránky, které se odkazují na velké množství jiných stránek
- Authority stránky
 - Stránky, na které se někdo často odkazuje

PageRank #1

- Hodnotí stránku dle její popularity
- Intuice
 - Linky jsou podobné jako citace v literatuře
 - U často citované stránky je očekáváno, že bude užitečnější
- PageRank počítá citace, ale své počítání rozšiřuje o
 - Váhu citací
 - Vyhlcení počtu citací
 - U každé stránky se uvažuje nenulový počet citací
- PageRank si lze představit jako výpočet pravděpodobnosti náhodného surfování

PageRank #2

- Data jsou reprezentována orientovaným grafem
 - Uzly – stránky
 - Hrany – linky
- Pro graf vytvoříme matici přechodů s pravděpodobnostmi přechodu do jiného uzlu

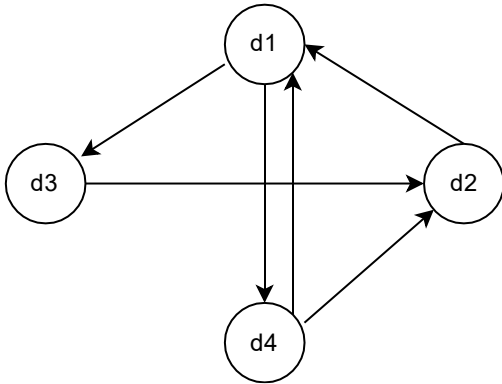


$$M = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

$$\sum_{j=1}^N M_{ij} = 1$$

M_{ij} = pravděpodobnost přechodu z d_i do d_j

PageRank #3

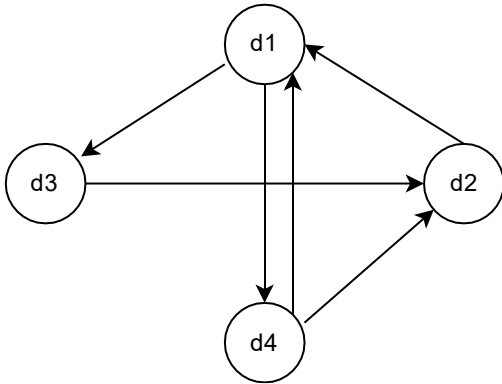


$$M = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

- Model s parametrem α
 - S pravděpodobností α skočíme náhodně na jinou stránku
 - S pravděpodobností $(1 - \alpha)$ náhodně vybereme link na současné stránce
- $p(d_i)$: PageRank skóre pro d_i = průměrná pravděpodobnost návštěvy stránky d_i

$$p_{t+1}(d_j) = (1 - \alpha) \sum_{i=1}^N M_{ij} p_t(d_i) + \alpha \frac{1}{N} \sum_{i=1}^N p_t(d_i)$$

PageRank #4



$$M = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

$$p_{t+1}(d_j) = (1 - \alpha) \sum_{i=1}^N M_{ij} p_t(d_i) + \alpha \frac{1}{N} \sum_{i=1}^N p_t(d_i)$$

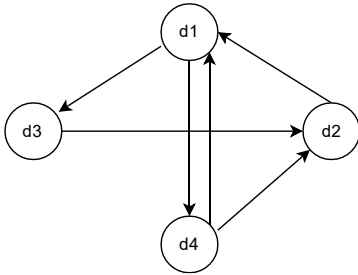


Vynechání časového indexu

$$p(d_j) = \sum_{i=1}^N \left[\frac{1}{N} \alpha + (1 - \alpha) M_{ij} \right] p(d_i) \quad \longrightarrow \quad \vec{p} = ((1 - \alpha)M + \frac{\alpha}{N}E)^T \vec{p}$$

$$E_{ij} = 1$$

PageRank #5



$$\vec{p} = ((1 - \alpha)M + \frac{\alpha}{N}E)^T \vec{p}$$

$$\alpha = 0,2 \quad N = 4$$

$$A = (1 - 0,2)M + 0,2 \frac{E}{N} = 0,8 \times \begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix} + 0,2 \times \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

$$\begin{bmatrix} p^{n+1}(d_1) \\ p^{n+1}(d_2) \\ p^{n+1}(d_3) \\ p^{n+1}(d_4) \end{bmatrix} = A^T \begin{bmatrix} p^n(d_1) \\ p^n(d_2) \\ p^n(d_3) \\ p^n(d_4) \end{bmatrix} = \begin{bmatrix} 0,05 & 0,85 & 0,05 & 0,45 \\ 0,05 & 0,05 & 0,85 & 0,45 \\ 0,45 & 0,05 & 0,05 & 0,05 \\ 0,45 & 0,05 & 0,05 & 0,05 \end{bmatrix} \times \begin{bmatrix} p^n(d_1) \\ p^n(d_2) \\ p^n(d_3) \\ p^n(d_4) \end{bmatrix}$$

- Počáteční hodnota $p^0(d_i) = 1/N$
- Iterujeme do konvergence
- Lze vypočítat PageRank pouze pro některé stránky

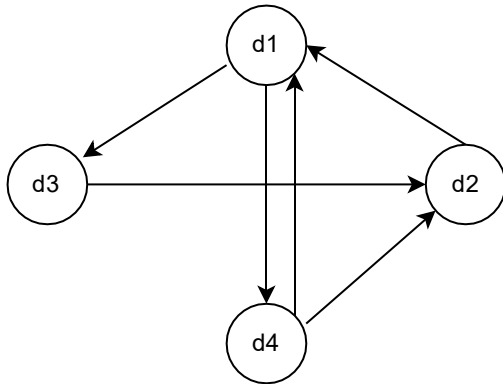
PageRank #6

- **PageRank udává důležitost stránky dle její popularity**
- Rychlý výpočet
- Problém se stránkou, která má 0 výstupních hran (0 linků)
 - Suma pravděpodobností nemůže být 1
 - Řešením je úprava parametru α
 - $\alpha = 1$, pro stránky s 0 linky
- Velké množství rozšíření
 - Např. topic-specific PageRank

HITS #1

- Hypertext-Induced Topic Search (HITS)
- PageRank nijak nerozlišuje autority a huby
 - Stránky hodnotí pouze dle jejich autority
- Intuice
 - Často citované stránky mají dobrou autoritu
 - Stránky, které citují hodně jiných stránek jsou dobré huby
- Hlavní myšlenka
 - Dobré autority budou citovány dobrými huby
 - Dobré huby odkazují na dobré autority
 - Např. odkazem na dobrou autoritu zvyšujeme své hub skóre

HITS #2



$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

A - Matice sousednosti

Počáteční hodnoty $h(d_i) = a(d_i) = 1$

Iterační proces

$$h(d_i) = \sum_{d_j \in OUT(d_i)} a(d_j)$$

$$a(d_i) = \sum_{d_j \in IN(d_i)} h(d_j)$$

Vektorizace



$$\vec{h} = A\vec{a}$$

$$\vec{a} = A^T\vec{h}$$

Po každé iteraci je potřeba normalizace

HITS vs PageRank

- HITS
 - Pro výsledné hodnocení stránky získáme 2 hodnoty
 - Pro získání skóre využívá pouze sousedy v grafu
 - Závislý na dotazu
 - První by měl získat relevantní stránky a poté začít hodnotit
 - Pomalejší než PageRank
- PageRank
 - Pro výsledné hodnocení stránky máme 1 hodnotu
 - Pro získání skóre využívá celý web
 - Nezávislý na dotazu
 - Velké množství rozšíření (např. závislost na dotazu)
 - Ve spojení s NN dosahuje lepší úspěšnosti než HITS

PageRank rozšíření

- Topic-Sensitive PageRank
- Query-Dependent PageRank
- Age-Based PageRank
 - Pomáhá nově vytvořeným stránkám dostat vyšší PageRank hodnocení
- Dirichlet PageRank
 - Stránky s malým počtem odkazů mají zvýšenou pravděpodobnost náhodného skoku
- Adaptive PageRank
 - Menší výpočetní náročnost
 - Dříve zastavíme iterace u stránek s malým PageRankem
 - Stránky s velkým PageRankem naopak iterujeme déle

Další algoritmy

- BlackRank
 - Model náhodného procházení více přibližuje k reálnému uživateli
 - Počítá i s návratem zpět na předchozí stránku
- WordRank
 - Odkazy mezi stránkami jsou váženy podobností jejich obsahu
 - Odkaz na podobnou stránku bude mít větší váhu
- FlexiRank
 - Spolupráce s uživatelem
 - Stránky jsou klasifikovány do kategorií (hlavní stránka, reklama, návod, ...)
 - Nejdříve se hledají relevantní dokumenty a poté je uživatel dotázán o jakou kategorii má zájem
- Velké množství dalších algoritmů (viz užitečná literatura)

Hodnocení pomocí ML

- Potřebujeme kombinovat více příznaků pro vyhledávání
 - Relevance dokumentu (např. výstup BM25)
 - PageRank
 - Vytvoření dalších (vlastních) příznaků
 - TF
 - Obsahuje url něco z dotazu?
 - ...
- Aplikace ML algoritmů, které na základě poskytnutých příznaků vyhodnotí výsledné skóre relevance

Užitečná literatura / kurzy

- Coursera kurz [Text Retrieval and Search Engines](#)
- [Content and link-structure perspective of ranking webpages: A review](#)
 - 2020 – aktuální přehled algoritmů pro vyhledávání na webu