Manažment znalostí (4)

OBSAH PREDNÁŠKY

- Meranie efektívnosti vyhľadávania
  - Porovnanie krivky presnosť-návratnosť pre 3 rôzne prístupy k IR
  - Sumarizačné mierky efektívnosti vyhľadávania
- Vyhľadávanie na webe
  - Veľmi stručná história vyhľadávania na webe
  - Architektúra crawler-indexer
  - Typy používateľských dopytov, kategórie, znalostný graf
  - Prezentácia výsledkov vyhľadávania - sumáre
  - Marketing založený na vyhľadávaní, spôsob fungovania Google Ads
- Vyhľadávanie s využitím štruktúry liniek
  - algoritmus PageRank
  - algoritmus HITS

Hodnotenie efektívnosti vyhľadávania pre **neusporiadanú množinu výsledkov**

- Uvažujme celú množinu výsledkov IR naraz (alebo aj ***unranked retrieval set***), pričom:
  - *q* je daný dopyt reprezentujúci informačnú potrebu
  - *R* je množina relevantných dokumentov ku *q*
  - $|R|$ je počet relevantných dokumentov ku *q*
  - *A* je množina dokumentov, ktoré vyhľadávací systém používajúci stratégiu *S* vráti ako odpoveď na *q*
  - $|A|$ je počet dokumentov vrátených *S* ako odpoveď na *q*
  - R_A je prienik množín *R* a *A*

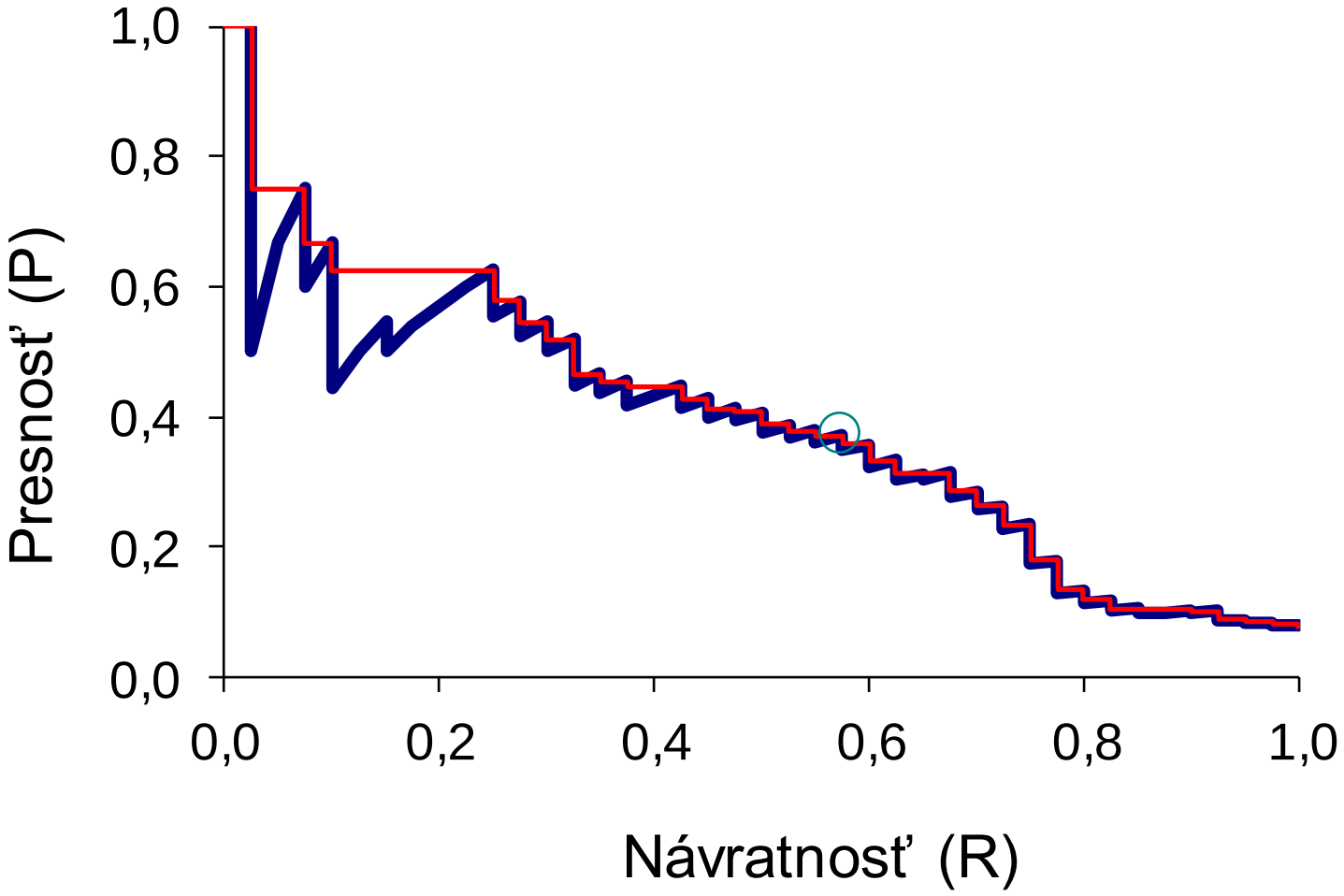
Návratnosť (recall) $N = \frac{|R_A|}{|R|}$

Presnosť (precision) $P = \frac{|R_A|}{|A|}$

Hodnotenie efektívnosti vyhľadávania pre **usporiadanú množinu výsledkov**

- Používateľ obyčajne nevidí celú množinu *A* (odpoveď na svoj dopyt *q*) naraz, ale postupne, dokumenty sú usporiadané podľa stupňa relevancie (***ranked retrieval set***)
- Teda **návratnosť a presnosť sa z pohľadu používateľa postupne menia**
- Priebeh presnosti, ako funkcie závislej od návratnosti sa zvykne zobrazovať graficky → tzv. **krivka presnosť – návratnosť**

Krivka presnosť – návratnosť



Vyhľadávanie založené na ontológii

Výsledky nášho výskumu publikovaného v: Jan Paralic, Ivan Kostial: **Ontology-based information retrieval**. Proceedings of the 14th International Conference on Information and Intelligent systems (IIS 2003), Varazdin, Croatia, p. 23-28 (*101 citácií na GoogleScholar*)

1. Vezme sa množina konceptov pre daný dopyt (Q_{con})
2. Vezme sa množina konceptov asociovaných s daným dokumentom ($D_{i,con}$)
3. Tieto dve množiny sa porovnajú nasledovnou mierkou podobnosti daného dokumentu $\mathbf{D_i}$ a dopytu $\mathbf{Q}$:

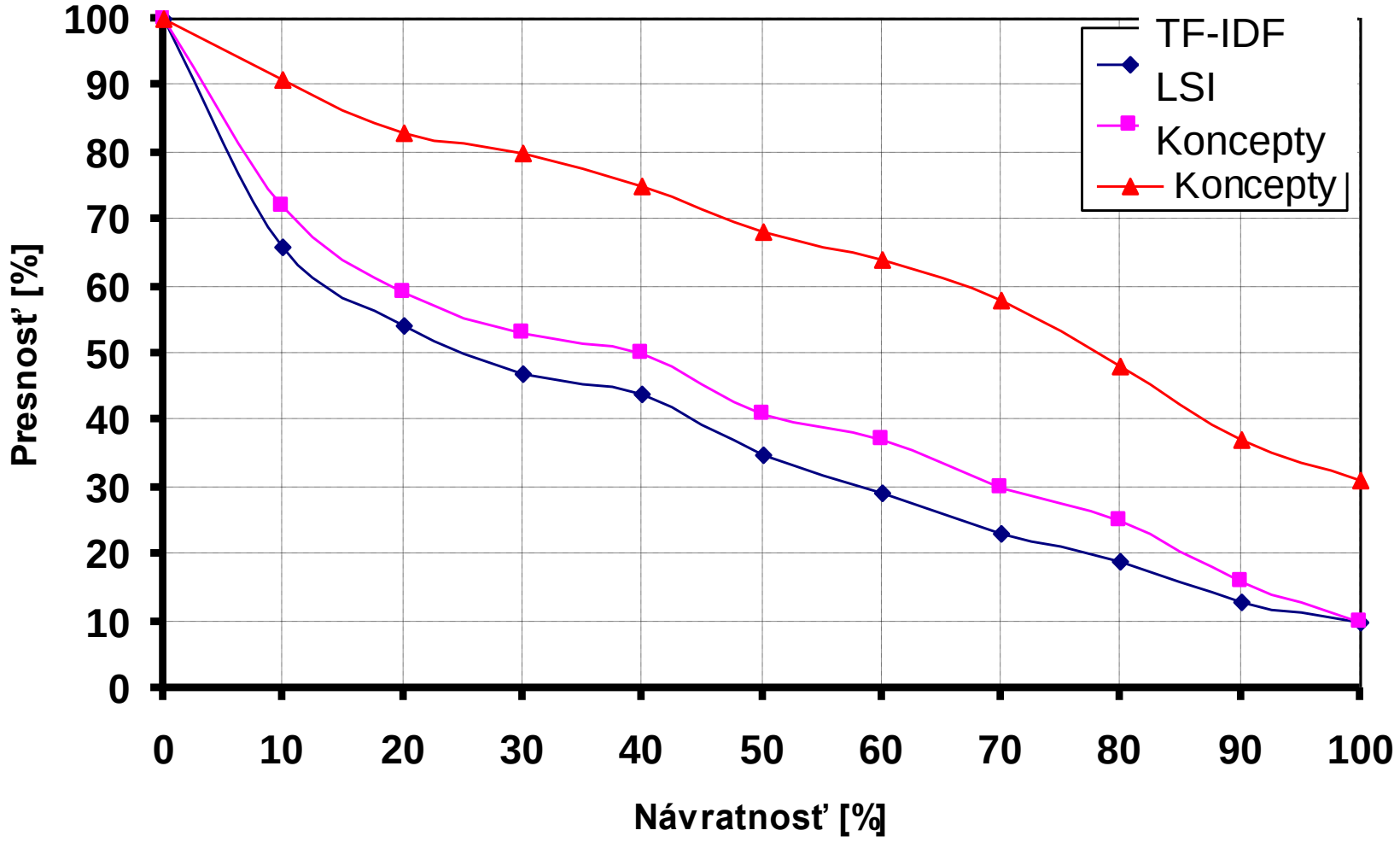
$$sim_{onto}(\mathbf{Q}, \mathbf{D_i}) = \begin{cases} \left|Q_{con} \cap D_{i,con}\right| & if \quad \left|Q_{con} \cap D_{i,con}\right| \neq 0 \\ k(=0,1) \end{cases}$$

4. Výsledná podobnosť sa vypočíta ako súčin podobnosti založenej na ontológii a podobnosti vypočítanej podľa vektorového modelu, resp. pomocou LSI modelu: $sim(\mathbf{Q}, \mathbf{D_i}) = sim_{onto}(\mathbf{Q}, \mathbf{D_i}) * sim_{TF-IDF}(\mathbf{Q}, \mathbf{D_i})$

Príklad – Použitá kolekcia dokumentov

- Kolekcia nazvaná ***Cystická fibróza*** (získaná z databázy MEDLINE)
  - Kolekcia pozostáva z 1239 dokumentov
  - Minimálna veľkosť dokumentu 0.12 kb, maximálna veľkosť 3.8 kb a priemerná veľkosť 1.045 kb
  - Ku kolekcii existuje aj súbor so 100 dopytmi
  - Pre každý dopyt je známa množina relevantných dokumentov
  - Každý dokument v odpovedi je ohodnotený číslom 0 až 8 (4 nezávislí experti hodnotili mieru relevancie 0 až 2)
  - Existuje 821 konceptov a priemerný počet konceptov priradených dokumentu je 2.8
  - Priemerný počet dokumentov asociovaných s jedným konceptom je 4.2

Porovnanie výsledkov pre 3 rôzne prístupy k vyhľadávaniu



Sumarizačné mierky efektívnosti vyhľadávania (1)

1. **Priemerná presnosť pri nájdených relevantných dokumentoch** (MAP – *mean average precision*)
   - Táto mierka favorizuje vyhľadávacie stratégie, ktoré rýchlo nájdu relevantné dokumenty

$$\overline{P}_{q_1} = \frac{1 + 0.66 + 0.5 + 0.4 + 0.3}{5} = 0.57 \qquad \overline{P}_{q_2} = \frac{0.33 + 0.25 + 0.2}{3} = 0.26$$

2. ***R*-presnosť** (*RP*) je presnosť vyhľadávacej stratégie *S* na |*R*|-tej pozícii, t.j. pri |*R*|-tom vrátenom dokumente

$$RP_{q_1} = \frac{4}{10} = 0.4 \qquad RP_{q_2} = \frac{1}{3} = 0.33$$

   - Táto mierka vlastne nie je sumarizačnou, popisuje iba jeden bod krivky presnosť – návratnosť, prax však ukazuje, že je silne korelovaná s *MAP*

Sumarizačné mierky efektívnosti vyhľadávania (2)

3. **Presnostné histogramy** sa používajú na detailnejšie porovnanie presnosti dvoch stratégií vyhľadávania (S_1 a S_2) pre viaceré dopyty $i = 1 .. N_q$

$$RP_{S_1/S_2}(i) = RP_{S_1}(i) - RP_{S_2}(i)$$

4. **Štatistiky v sumarizačnej tabuľke** – napr. počet otázok, celkový počet vrátených dokumentov, z nich celkový počet relevantných dokumentov, a pod.

Používateľsky orientované mierky efektívnosti vyhľadávania

- U je podmnožina R takých dokumentov, ktoré sú používateľovi už známe
- $R_k = A \cap U$ je množina používateľovi známych dokumentov v odpovedi A
- R_U je množina relevantných dokumentov v odpovedi A, ktoré používateľovi neboli predtým známe

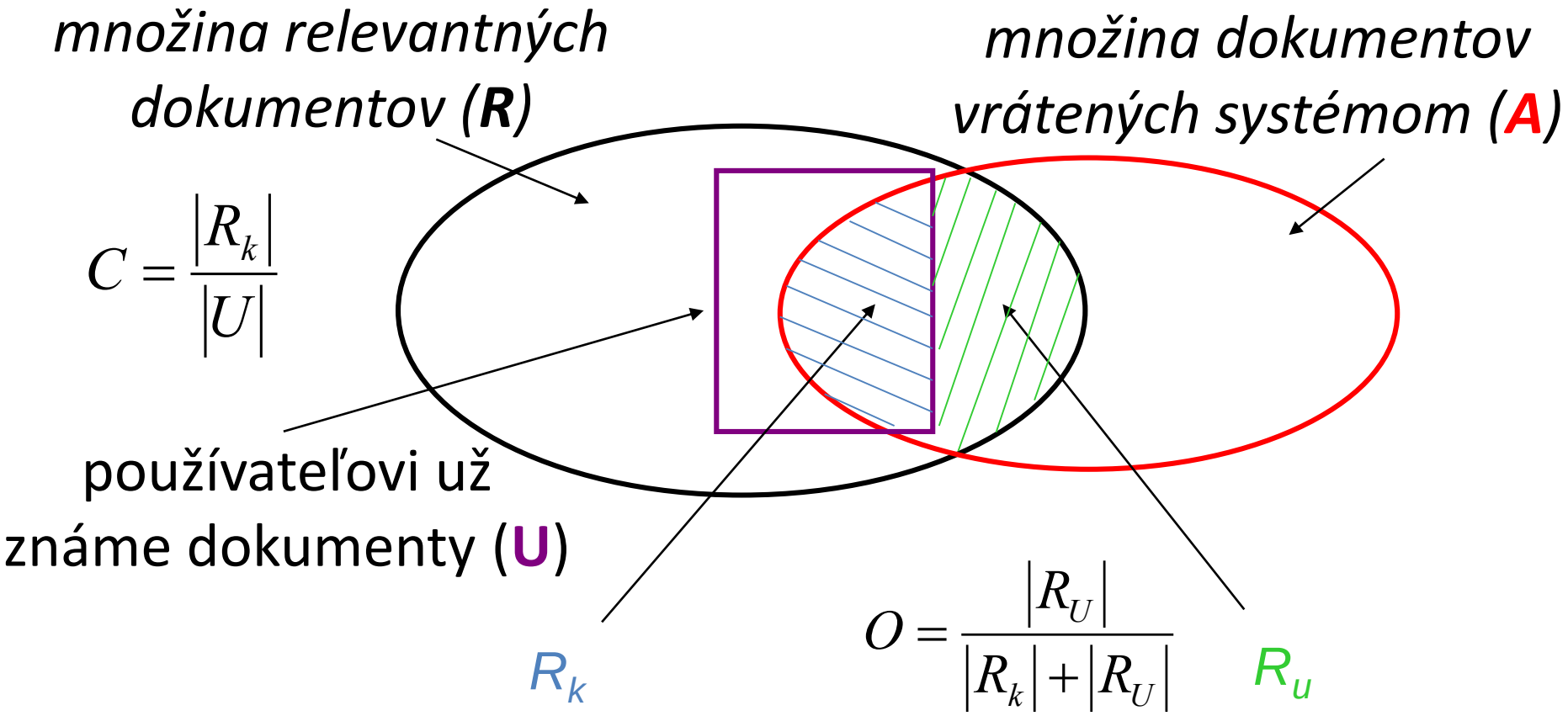
5. **Pokrytie (coverage) *C*** je definované nasledovne:

$$C = \frac{|R_k|}{|U|}$$

6. **Novosť (novelty) *O*** je definovaná nasledovne:

$$O = \frac{|R_U|}{|R_k| + |R_U|} = \frac{|R_U|}{|R_A|}$$

Vysvetlenie významu množín pri používateľsky definovaných mierkach efektívnosti vyhľadávania



Vylepšovanie bežiaceho IR systému

- Používateľské štúdie sú dobrý nástroj, najmä v čase návrhu, ale sú časovo náročné a nákladné
- Pre bežiaci IR systém sa najčastejšie používa metóda zvaná **A/B test**:
  - Pre takýto test sa spraví práve jedna zmena (systém B) aktuálneho systému (systém A), ktorej vplyv chceme ohodnotiť
  - Časť používateľských požiadaviek (1 až 10%) sa presmeruje na zmenený systém B, zvyšné spracúva aktuálne bežiaci systém A
  - Porovnajú sa sledované parametre (napr. na čo ľudia kliknú, frekvencia klikaní na prvý odkaz v zozname, koľko bolo odoslaných dopytov, či boli dopyty opustené, ako dlho trvalo, kým používatelia klikli na výsledok atď.) medzi systémami A a B
  - Pri dostatočne veľkom počte používateľov možno takýmto spôsobom lacno a rýchle overiť vplyv navrhovanej zmeny

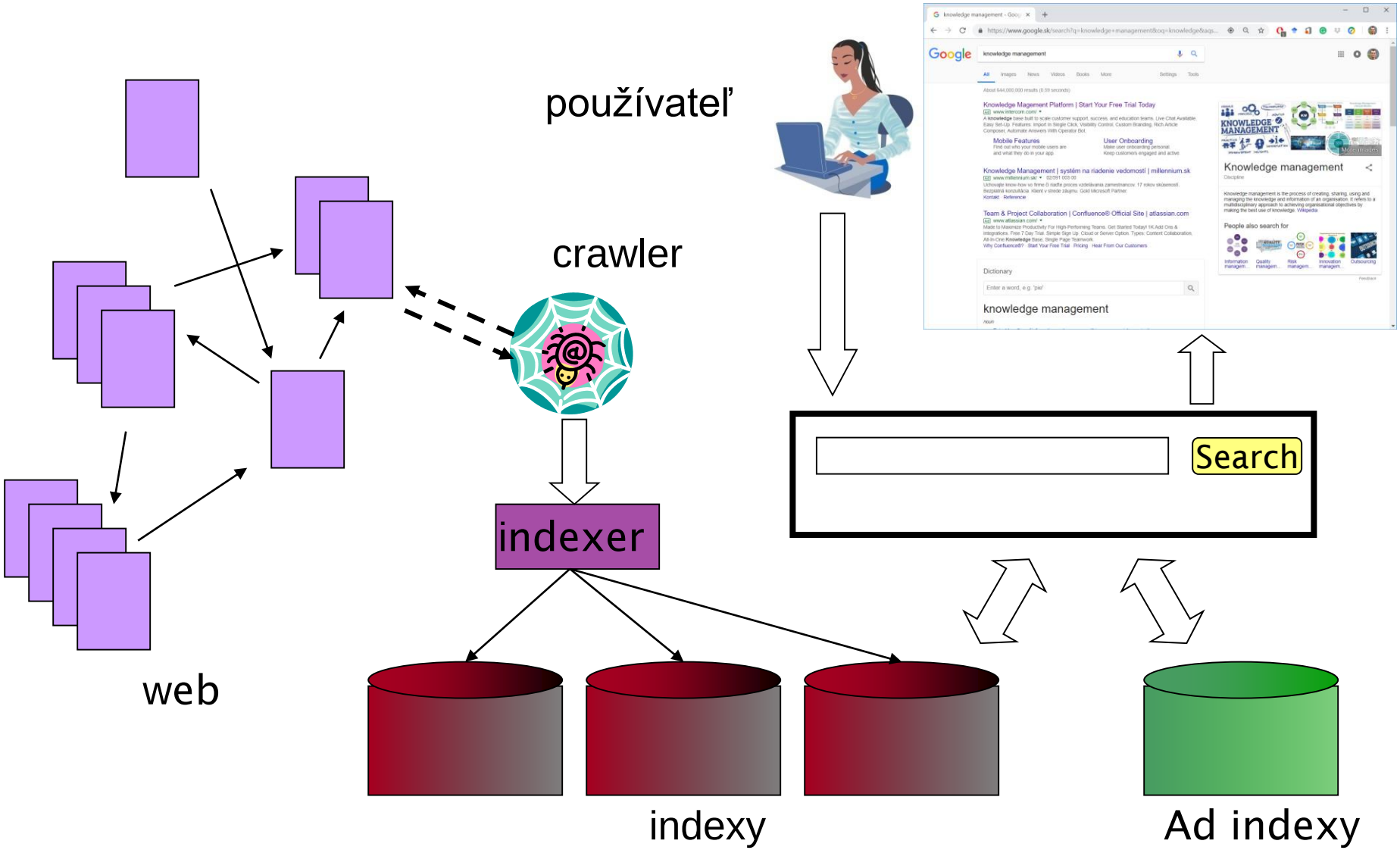
Vyhľadávanie na webe – veľmi stručná história

- Vyhľadávače využívajúce iba plno-textové indexovanie webových stránok (obdobie 1994 – 1997, napr. Infoseek od roku 1994, Altavista a Excite od roku 1995)
- Kategórie subjektov napĺňané odkazmi na webové stránky, napr. *Yahoo! Directory 1994-2014,* neskôr DMOZ, resp. *Open Directory Project 1998-2017*, resp. *Google directory (do 2011*)
- Využitie štruktúry liniek pri vyhľadávaní (od roku 1998, najskôr *Google*)
- Sponzorované vyhľadávanie (alebo aj search marketing fungujúci na princípe „pay per click“ – od r. 1998 *Yahoo!,* neskôr aj *Google AdWords od roku 2000*)
- Sémantické vyhľadávanie – napr. znalostný graf (2012)
- Odpovedanie na otázky – napr. WolframAlpha, alebo IBM Watson, nárast aplikácií využívajúcich chatboty

Vyhľadávacie stroje

- Väčšina vyhľadávacích strojov má centralizovanú architektúru typu „**crawler-indexer**“
- „**Crawlers**“ sú programy ktorých cieľom je čo najrýchlejšie a najefektívnejšie získať stránky z webu, vrátane štruktúry ich prepojení
- **Indexer** má za úlohu indexáciu web stránok získaných a parsovaných crawlerom a ich uloženie vo vhodnej štruktúre (indexe)
- Môžete si pozrieť ako funguje vyhľadávanie v Google alebo pozrite si video

Architektúra „crawler-indexer“



Algoritmy vyhľadávania

- Algoritmy vyhľadávania posudzujú mnoho faktorov vrátane:
  - slov v dopyte, relevancie
  - použiteľnosti stránok
  - odbornosti zdrojov
  - polohy vyhľadávajúceho a jeho nastavení
- Váha jednotlivých faktorov závisí od **typu dopytu**
  - napr. čerstvosť obsahu má napríklad väčší význam pri odpovedaní na dopyty týkajúce sa aktuálnych spravodajských tém ako v prípade definícií zo slovníka

Typy používateľských dopytov

- Tri veľké skupiny:
  1. **Navigačné dopyty** (*15% – 25%*) – používateľ hľadá domovskú web stránku nejakej entity („*TU v Košiciach*“). Používateľ v tomto prípade očakáva hľadanú odpoveď na prvom mieste vo výsledkoch.
     - *V prípade zadania adresy priamo jej zobrazenie na mape*
  2. **Transakčné dopyty** (*25% – 35%*) – predchádzajú realizácii nejakej transakcie na webe (*kúpa produktu, stiahnutie súboru, rezervácia* a pod.). V tomto prípade je potrebné vrátiť zoznam služieb, ktoré poskytujú rozhranie pre žiadané transakcie.
  3. **Informačné dopyty** (*45% – 60%*) – používateľ hľadá všeobecné informácie na určitú oblasť („*knowledge management*“). Typicky neexistuje jedna web stránka, ktorá by obsahovala všetky potrebné informácie.
     - *Podrobnejšie informácie dostupné pre entity znalostného grafu*
     - *Prípadne priamo odpoveď na priamu otázku (featured snippet)*

Table 3.2 Classification rules according to (Jansen et al., 2008)

Navigational
queries containing company/business/organization/people names
queries containing domains suffixes
queries containing parts of URL address
queries length (i.e., number of terms in query) less than 3
searcher viewing the first search engine results page
Transactional
queries containing terms related to movies, songs, lyrics, recipes, images, humor, etc.
queries with 'obtaining' terms (e.g. lyrics, recipes, etc.);
queries with 'download' terms (e.g. download, software, etc.);
queries relating to image, audio, or video collections;
queries with 'audio', 'images', or 'video' as the source;
queries with 'entertainment' terms (pictures, games, etc.);
queries with 'interact' terms (e.g. buy, chat, etc.);
queries with movies, songs, lyrics, images, and multimedia or compression file extensions (jpeg, zip, etc.).
Informational
uses question words (i.e., 'ways to', 'how to', 'what is', etc.);
queries with natural language terms;
queries containing informational terms (e.g. list, playlist, etc.);
queries that were beyond the first query submitted;
queries where the searcher viewed multiple results pages;
queries length (i.e., number of terms in a query) greater than 2;
queries that do not meet criteria for navigational or transactional.

Personalizácia vyhľadávania

- Existuje niekoľko aspektov tejto úlohy:
  - **Personalizačná stratégia**
    - **Úprava dopytu** (rozšírenie alebo preformulovanie)
    - **Preusporiadanie výsledkov vyhľadávania**
  - **Zdroj dát pre personalizačné rozhodovanie**:
    - **Založené na obsahu** (podobnosť medzi dokumentmi)
    - **Kolaboratívna filtrácia** (podobnosť medzi používateľmi)
    - **Hybridné** (kombinácia oboch prístupov)
  - **Časový interval, z ktorého sa berú dáta pre personalizáciu**:
    - **Dlhodobé** (celá história interakcií používateľa s vyhľadávačom)
    - **Krátkodobé** (uvažuje iba posledné interakcie)

Prezentácia výsledkov vyhľadávania

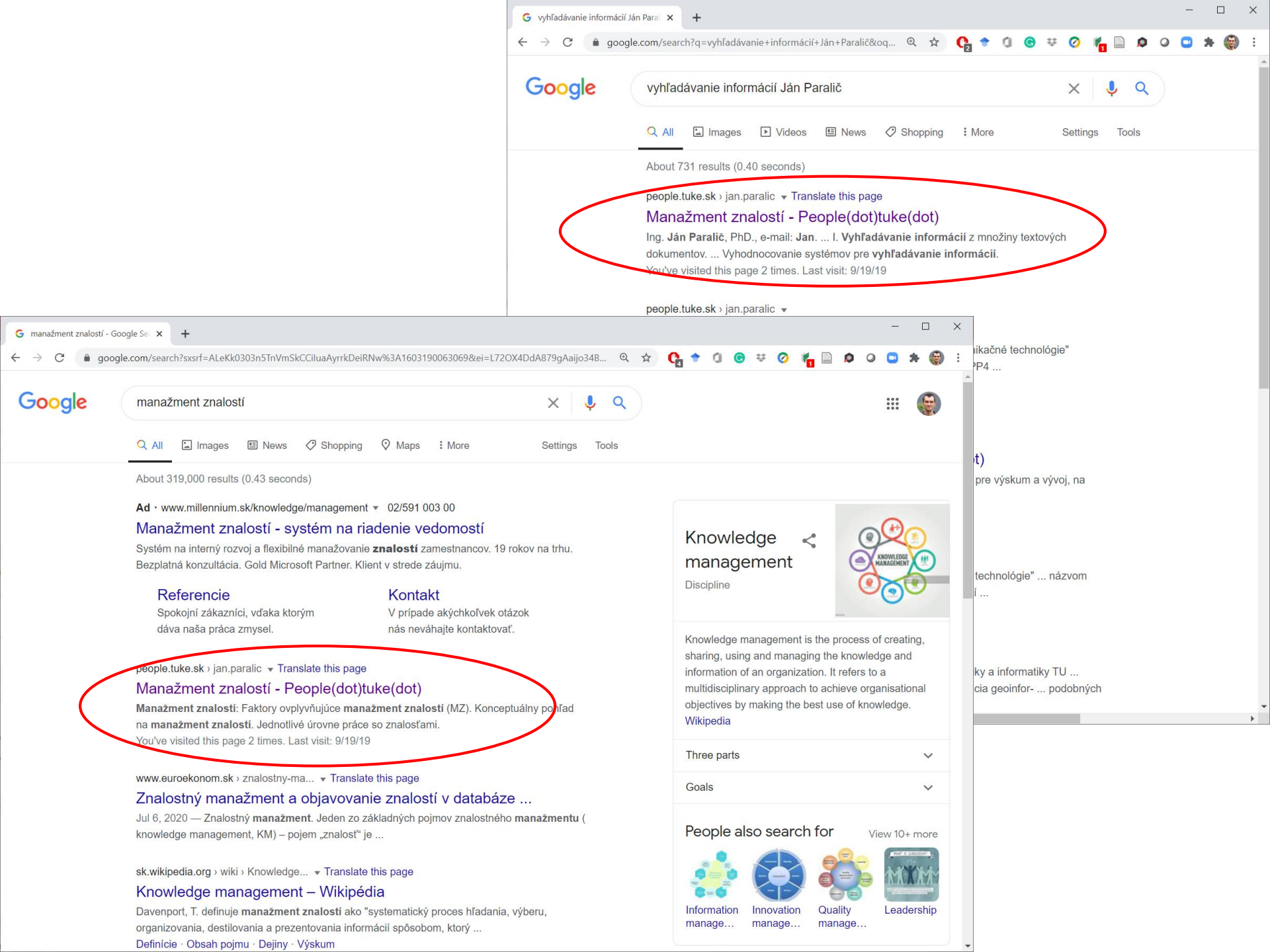
- Systém IR vráti usporiadaný zoznam dokumentov (podľa miery relevancie)
- Preddefinovaný počet dokumentov s krátkym popisom – **sumárom (snippet)**
  - Obsah sumáru je dôležitý – viac o sumároch pozri ďalej
- Platené odkazy na zadaný dopyt (Ads)
- Ale môžu to byť aj ďalšie užitočné odpovede:
  - Napr. definícia z wikipédie, resp. uzol zo znalostného grafu, najnovšie už aj priamu odpoveď na zadanú otázku (featured snippet)

Sumáre vo výsledkoch IR systémov (1)

- Dva základné druhy sumárov:
  - **statické** – nezávislé na dopyte ktorý viedol k vyhľadaniu daného dokumentu, stále rovnaký
  - **dynamické** – prispôsobené konkrétnemu dopytu, snažia sa ukázať, prečo bol daný dokument vybraný ako relevantný k dopytu
- **Statické sumáre** typicky predstavujú časť dokumentu
  - napr. prvých 50 slov, uložené do cache v čase indexácie
  - výber reprezentatívnej množiny viet z dokumentu - použitie NLP pre skórovanie viet a výber najlepších
  - sofistikované techniky sumarizácie textov – používané v experimentálnych systémoch

Sumáre vo výsledkoch IR systémov (2)

- **Dynamické sumáre** Prezentujú jedno alebo viac „okien“ v dokumente, ktoré obsahujú niekoľko termov z dopytu
  - vyžaduje rýchle vyhľadanie okien v cache pamäti dokumentov
  - Skórovanie nájdených okien vzhľadom na dopyt (príznaky ako veľkosť a poloha okna v dokumente)
- Aké sumáre používa vyhľadávač Google?


vyhľadávanie informácií Ján Paralič
google.com/search?q=vyhľadávanie+informácií+Ján+Paralič&oq...
Google
vyhľadávanie informácií Ján Paralič
All
Images
Videos
News
Shopping
More
Settings
Tools
About 731 results (0.40 seconds)
people.tuke.sk › jan.paralic
Translate this page
Manažment znalostí - People(dot)tuke(dot)
Ing. Ján Paralič, PhD., e-mail: Jan. ... I. Vyhľadávanie informácií z množiny textových dokumentov. ... Vyhodnocovanie systémov pre vyhľadávanie informácií.
You've visited this page 2 times. Last visit: 9/19/19
people.tuke.sk › jan.paralic
manažment znalostí - Google Se
google.com/search?sxsrf=ALeKk0303n5TnVmSkCCiluaAyrrkDeiRNw%3A1603190063069&ei=L72OX4DdA879gAaijo34B...
Google
manažment znalostí
All
Images
News
Shopping
Maps
More
Settings
Tools
About 319,000 results (0.43 seconds)
Ad · www.millennium.sk/knowledge/management
02/591 003 00
Manažment znalostí - systém na riadenie vedomostí
Systém na interný rozvoj a flexibilné manažovanie znalostí zamestnancov. 19 rokov na trhu. Bezplatná konzultácia. Gold Microsoft Partner. Klient v strede záujmu.
Referencie
Spokojní zákazníci, vďaka ktorým dáva naša práca zmysel.
Kontakt
V prípade akýchkoľvek otázok nás neváhajte kontaktovať.
people.tuke.sk › jan.paralic
Translate this page
Manažment znalostí - People(dot)tuke(dot)
Manažment znalostí: Faktory ovplyvňujúce manažment znalostí (MZ). Konceptuálny pohľad na manažment znalostí. Jednotlivé úrovne práce so znalosťami.
You've visited this page 2 times. Last visit: 9/19/19
www.euroekonom.sk › znalostny-ma...
Translate this page
Znalostný manažment a objavovanie znalostí v databáze ...
Jul 6, 2020 — Znalostný manažment. Jeden zo základných pojmov znalostného manažmentu (knowledge management, KM) – pojem „znalosť" je ...
sk.wikipedia.org › wiki › Knowledge...
Translate this page
Knowledge management – Wikipédia
Davenport, T. definuje manažment znalostí ako "systematický proces hľadania, výberu, organizovania, destilovania a prezentovania informácií spôsobom, ktorý ...
Definície · Obsah pojmu · Dejiny · Výskum
Knowledge management
Discipline
KNOWLEDGE MANAGEMENT
Knowledge management is the process of creating, sharing, using and managing the knowledge and information of an organization. It refers to a multidisciplinary approach to achieve organisational objectives by making the best use of knowledge.
Wikipedia
Three parts
Goals
People also search for
View 10+ more
Information manage…
Innovation manage…
Quality manage…
Leadership
ikačné technológie"
PP4 ...
t)
pre výskum a vývoj, na
technológie" ... názvom
í ...
ky a informatiky TU ...
cia geoinfor- ... podobných

Marketing založený na vyhľadávaní

- Marketingová stratégia, kde hlavným nástrojom je vyhľadávač.
- Cieľom je dosiahnuť, aby sa daná web stránka ocitla vo výsledkoch pri hľadaní určitých kľúčových slov čo najvyššie.
- To sa dá dosiahnuť:
  a) **Optimalizáciou web stránok pre vyhľadávače** (SEO) – v rámci stránky samotnej, alebo aj mimo nej
  b) **Využitím platených služieb** – platenie za reklamu založené na princípe CPC (*cost per click*) – napr. Google Ads, Microsoft Advertising ...

Ako funguje Google Ads (1)

- Google Ads
  - Reklama vo vyhľadávaní, Reklama v obsahovej sieti
  - Videoreklama, Reklama na aplikáciu
- Definovanie marketingových kampaní:
  - Názov, dátum ukončenia, denný rozpočet, **BID** (maximálne CPC)
  - Distribučné preferencie (Google homepage, sieť partnerských vyhľadávačov, obsahová sieť – partneri využívajúci AdSense)
  - Výber cieľových jazykov, geografické zacielenie,
  - Časové úseky dňa, kedy kampaň bude bežať
  - Miesto zobrazenia reklamy (s tým súvisí cena)
  - Demograficky závislé ponuky (len pre „social media sites“)
- Výsledok dopytu v sekcii sponzorovaných liniek závisí od:
  - **BID** (maximálna cena za klik, ktorú je ochotný zadávateľ zaplatiť)
  - **Quality Score** (skóre kvality)
  - Očakávaný vplyv dodatočných zdrojov k reklame (Ad extensions) a ďalších formátov reklamy (Ad formats)

Ako funguje Google Ads (2)

- Vypočíta sa odhad ceny CPC ponuky potrebnej pre dosiahnutie prvej pozície na daný dopyt
  - Čím vyššie skóre kvality, tým nižšia odhadovaná cena CPC
  - Tento model podporuje dobre navrhnuté marketingové kampane
- Usporiadanie marketingových ponúk pre dané kľúčové slovo dopytu podľa hodnoty „*Ad Rank*“

  Ad Rank = maximálne_CPC x skóre_kvality
- Víťazná ponuka sa zobrazí a zadávateľ reklamy za ňu zaplatí najnižšiu cenu, ktorá by mu ešte zabezpečila danú pozíciu
  - Nie je možné zaistiť si stopercentne prvé miesto v zobrazených výsledkoch sponzorovaných liniek
  - Systém vedie k optimalizácii kampaní takým spôsobom, aby dosiahli čo najvyššie skóre kvality

Nonprofit Marketing Immersion (Google Ad Grants)

- Študentské tímy (po 2-5 študentov), verifikované vysokoškolským učiteľom. Najprv absolvujú on-line Google marketing trainings zakončené skúškou.
- Študentské tímy dostanú AdGrants nonprofit pre vybranú neziskovú organizáciu, pre ktorú pripravia marketingovú kampaň v rámci prideleného voľného limitu až do 10.000 USD na mesiac
- Marketingová kampaň: 4 súvislé týždne
- Analýza kampane po jej ukončení
- Viac informácií na: https://www.google.com/grants/get-help/nonprofit-marketing-immersion/

Využitie štruktúry liniek

Algoritmus PageRank (1)

- Vznikol v rámci projektu na univerzite Stanford a znamenal začiatok Google
  - Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd: **The PageRank Citation Ranking: Bringing Order to the Web.** Technical Report. Stanford InfoLab. 1999.
- **Využíva štruktúru web liniek pre výpočet hodnotenia kvality** (**PageRank**) jednotlivých web stránok
- Každá web stránka má unikátny PageRank, nezávislý na dopyte, ale iba na štruktúre prepojení
- PageRank teda nevyjadruje relevanciu stránky vzhľadom na daný dopyt

Algoritmus PageRank (2)

A. Jeden pohľad: **PageRank** možno chápať **ako simuláciu náhodného používateľa**, ktorý:
   1. vychádza zo stránky na náhodnom URL,
   2. klikne náhodne na niektorú z liniek danej stránky,
   3. po chvíli (náhodnej dĺžky) sa začne nudiť a skočí na stránku s iným náhodným URL.

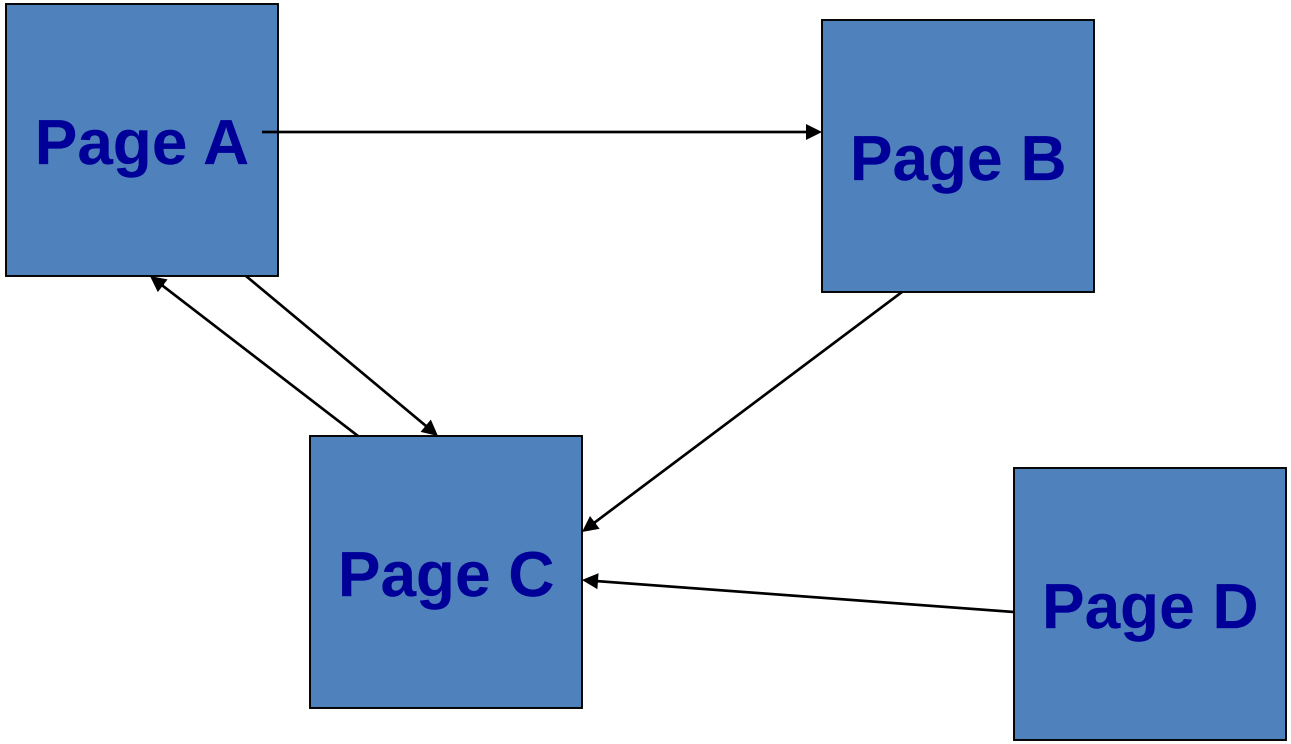
B. Druhý pohľad: **PageRank ako hlasovanie stránok medzi sebou**:
   - PageRank interpretuje linku zo stránky A na stránku B ako hlas, ktorý odovzdá stránka A stránke B
   - Naviac je tento hlas vážený významnosťou stránky A, ktorá dáva svoj hlas stránke B, t.j.
   - PageRank danej stránky sa zvyšuje tým viac, čím vyšší je PageRank stránok, ktoré sa na danú stránku odkazujú

Algoritmus PageRank (3)

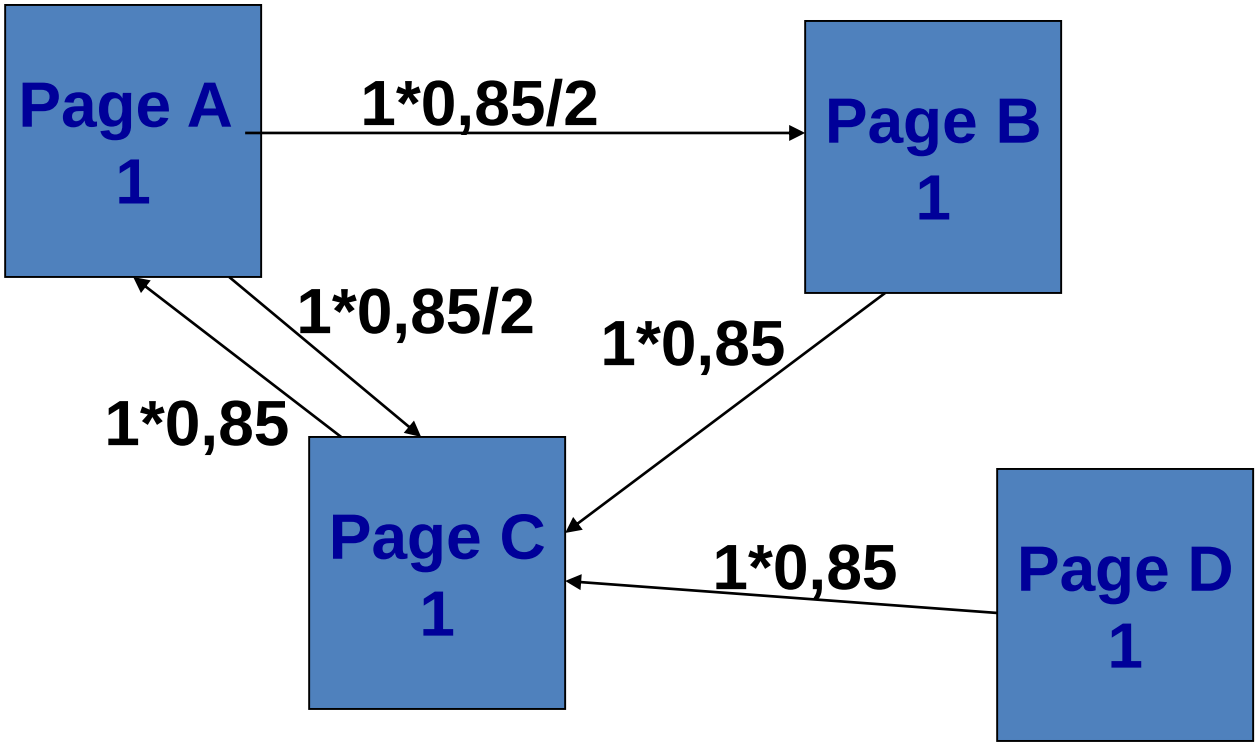
$$PR(A) = (1-d) + d*(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)})$$

- ***d***: tlmiaci faktor, nastavený na 0,85 – 0,9
- ***T_1***, …, ***T_n***: stránky odkazujúce sa na stránku *A*
- ***PR(A)***: PageRank stránky *A*
- ***PR(T_i)***: PageRank stránky T_i
- ***C(T_i)***: počet liniek vychádzajúcich zo stránky T_i

Príklad výpočtu ohodnotení PageRank



PageRank – 1. iterácia



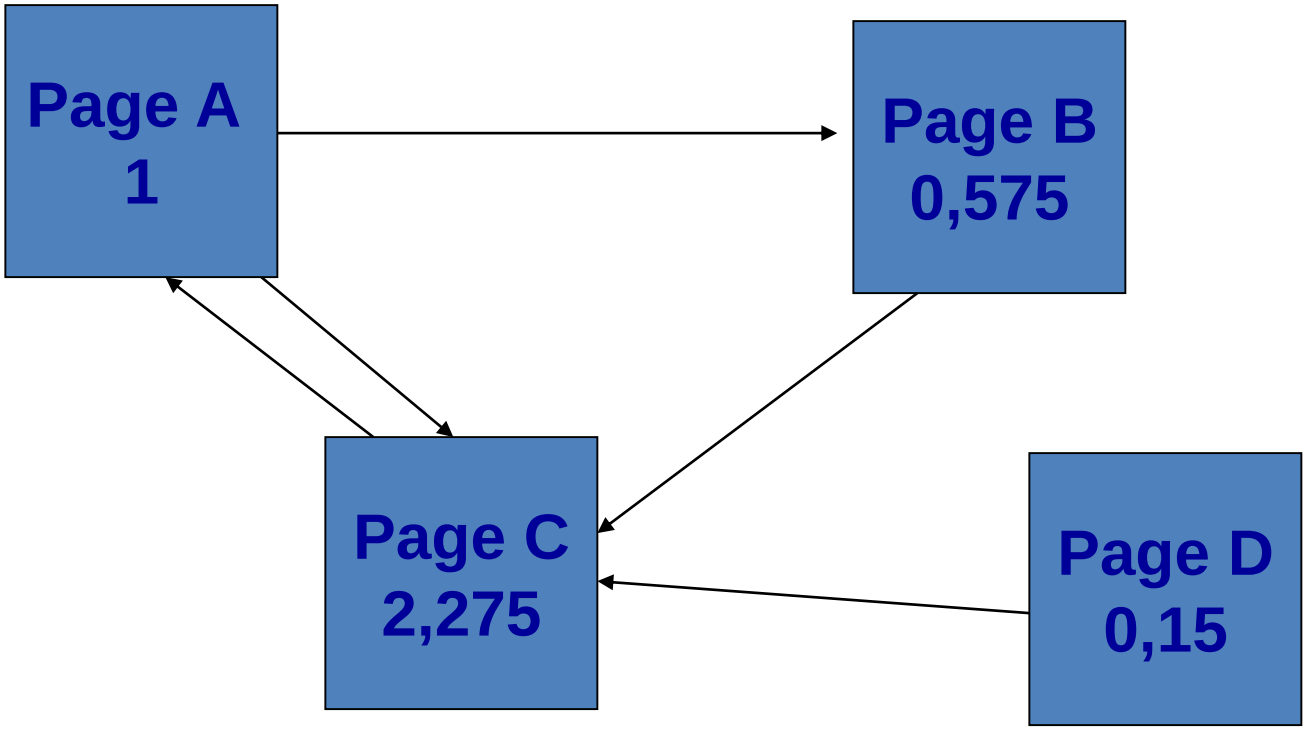
Page A: 0,85 (od stránky C) + 0,15 (nepresunuté) = **1**

Page B: 0,425 (od stránky A) + 0,15 (nepresunuté) = **0,575**

Page C: 0,85 (od stránky D) + 0,85 (od stránky B) + 0,425 (od stránky A) + 0,15 (nepresunuté) = **2,275**

Page D: nezíska žiadne, ale nepresunuté 0,15 = **0,15**

PageRank – 2. iterácia



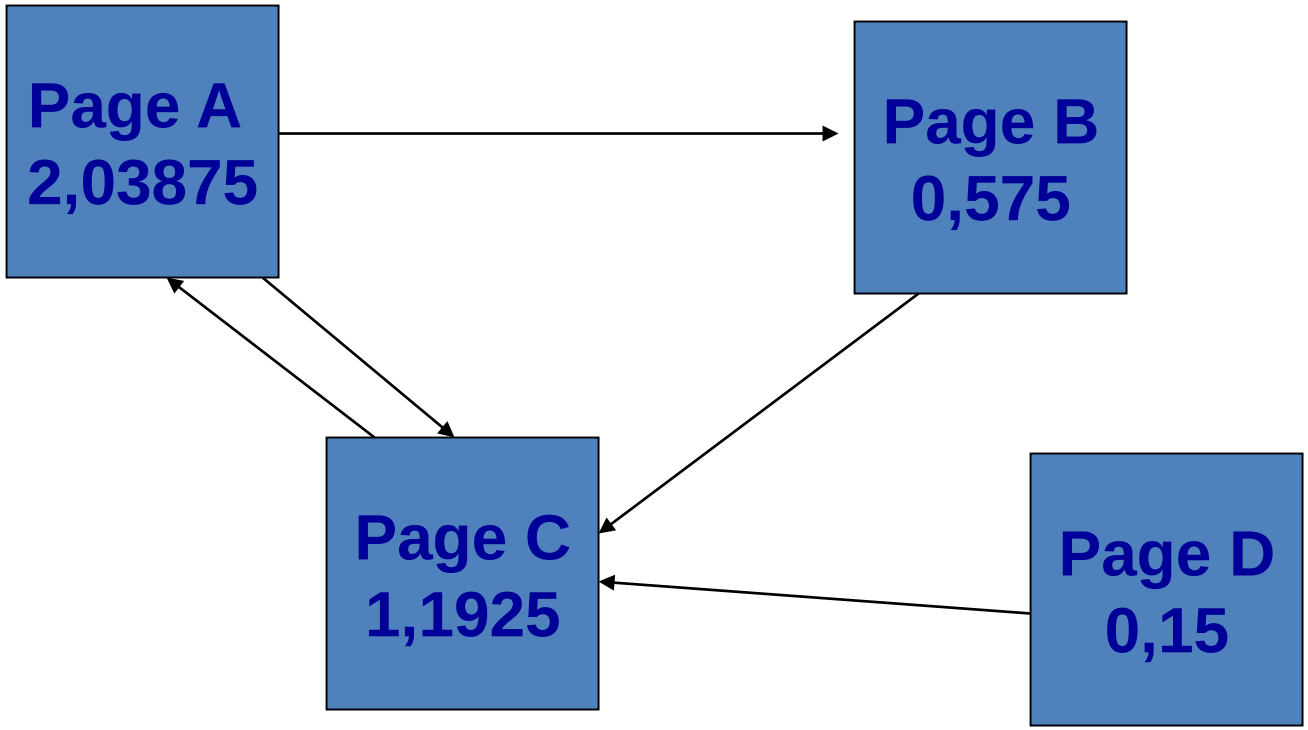
Page A: 2,275*0,85 (od stránky C) + 0,15 (nepresunuté) = **2,08375**

Page B: 1*0,85/2 (od stránky A) + 0,15 (nepresunuté) = **0,575**

Page C: 0,15*0.85 (od stránky D) + 0,575*0,85(od stránky B) + 1*0,85/2 (od stránky A) + 0,15 (nepresunuté) = **1,19125**

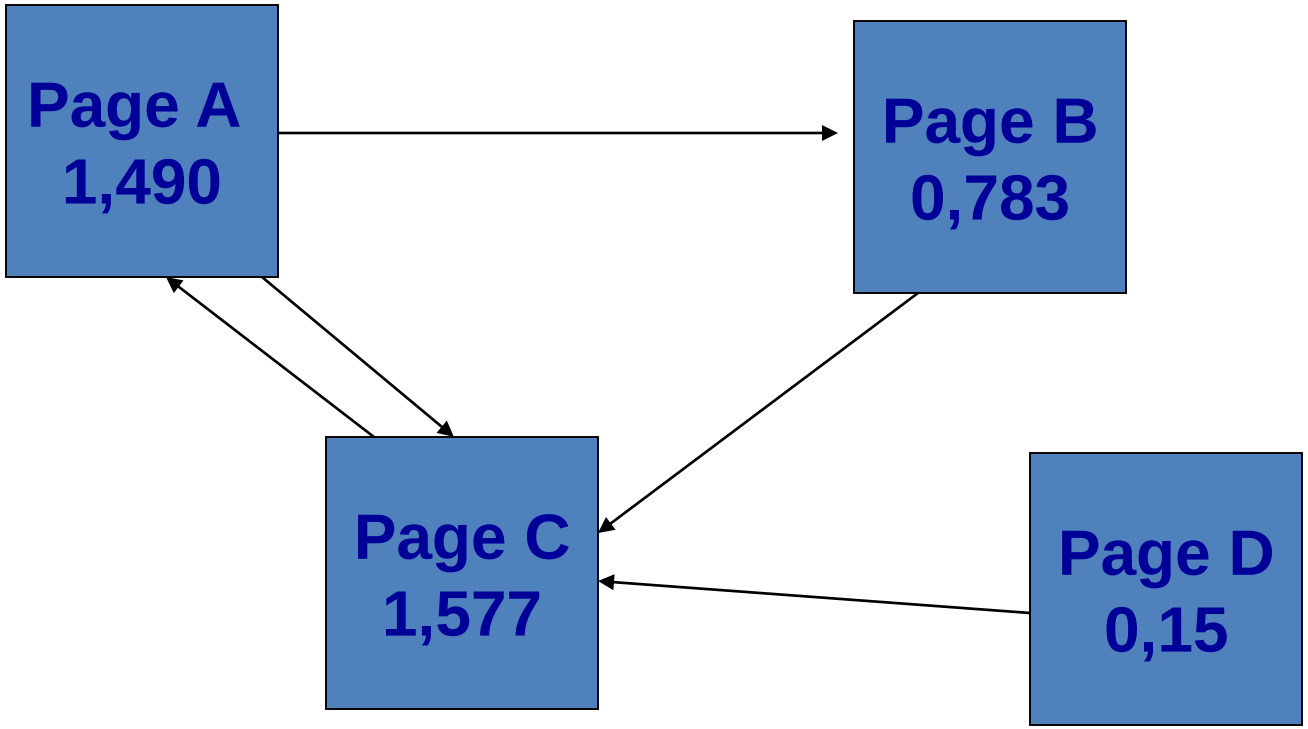
Page D: nezíska žiadne, ale nepresunuté 0,15 = **0,15**

PageRank – 3. iterácia



PageRank – 20. iterácia

- **Po 20 iteráciách** sa už hodnoty PageRank v podstate nemenia:



- Stránka C má teda najväčší význam v danej sieti, za ňou nasleduje stránka A
- Viac iterácií algoritmu vedie ku konvergencii váh PageRank

Usporiadanie výsledkov vo vyhľadávači Google

- Používa algoritmus PageRank ako jedno z kritérií pre usporiadanie výsledkov na dopyt používateľa
- Kritérií, ktoré sa berú do úvahy pri usporadúvaní výsledkov je viac ako dvesto, medzi nimi napr.:
  - Frekvencia termov
  - Blízkosť termov
  - Pozícia termov (názov, na začiatku stránky, a pod.)
  - Charakteristiky termu (hrubé písmo, kapitálky, a pod.)
  - **Informácia z analýzy liniek (*PageRank*)**
  - Informácia o kategóriách
  - ... a mnohé ďalšie

Algoritmus HITS (1)

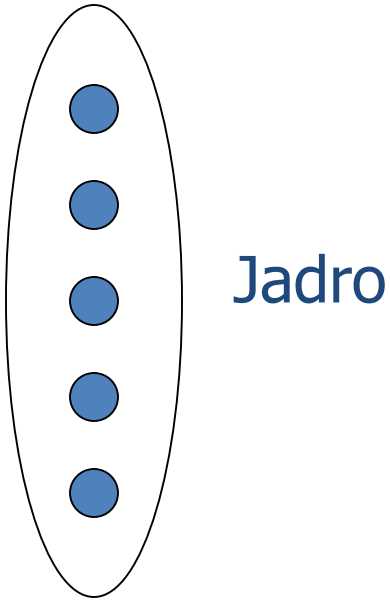
- **HITS** (Hypertext-Induced Topic Search) navrhnutý Jon Kleinbergom počas pobytu v IBM
- IBM rozšírila HITS v systéme Clever
- Clever nie je vyhľadávací stroj určený na prácu v reálnom čase
- Usporadúva stránky vzhľadom na ich relevanciu k používateľovmu dopytu (na rozdiel od PageRank)

Algoritmus HITS (2)

- Každá web stránka je hodnotená dvojicou váh:
  - **váha hub**
  - **váha autorita**
- Dobrý **hub** je web stránka, ktorá odkazuje na mnoho dobrých autorít
- Dobrá **autorita** je web stránka, na ktorú sa odkazuje mnoho dobrých hubov
- Výpočet váh autorít a hubov iteráciami konverguje k vlastným vektorom M^TM a MM^T, kde M je matica susednosti orientovaného podgrafu webu

Algoritmus HITS (3)

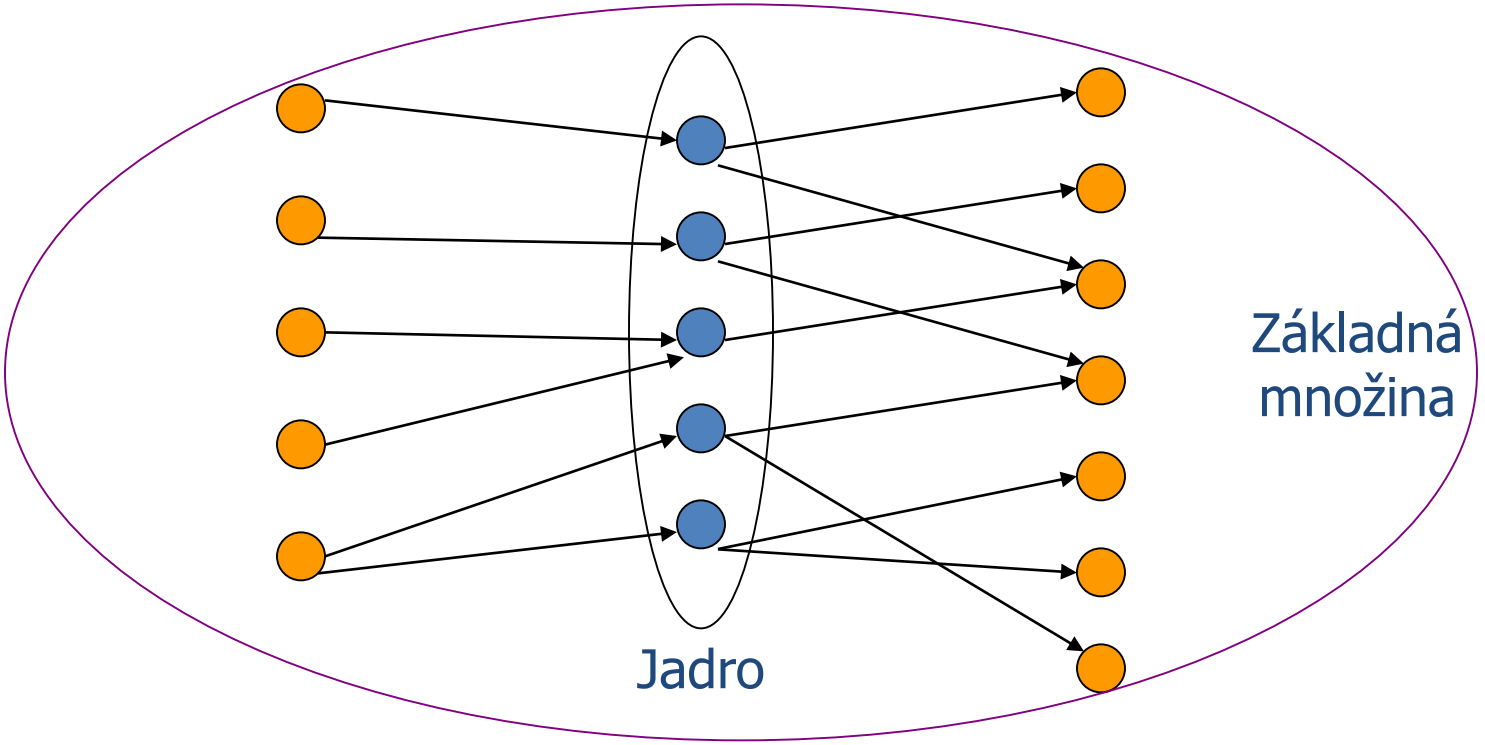
1. **Použitím dopytu sa získa** tzv. ***jadro (root set)*** **stránok** z textového vyhľadávacieho stroja



Algoritmus HITS (4)

2. **Jadro sa rozšíri na *základnú množinu (base set)* web stránok**
   - zahrnutím všetkých stránok na ktoré sa odkazujú stránky z jadra
   - a všetkých stránok ktoré sa odkazujú na nejakú stránku z jadra

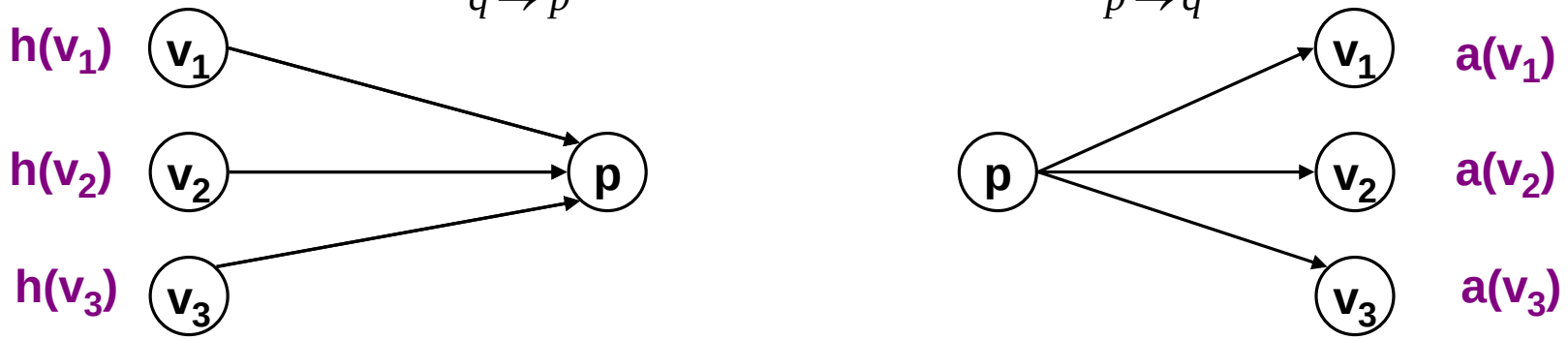
   Typická základná množina obsahuje *1000-5000* stránok



Algoritmus HITS (5)

3. **Algoritmus iteratívne počíta váhy autority *a(p)* a váhy hub *h(p)***
   - Nastav váhy *a(p)* = 1 a váhy hub *h(p)* = 1 pre všetky p
   - Opakuj nasledovné 2 operácie

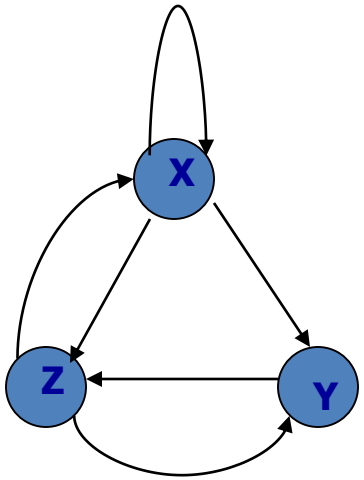
$$a(p) = \sum_{q \to p} h(q) \qquad h(p) = \sum_{p \to q} a(q)$$



4. **Nakoniec znormalizuj *a* a *h* na jednotkovú veľkosť**
5. **Výstupom** algoritmu **je zoznam najlepších autorít a niekoľko najlepších hubov**.

Príklad výpočtu algoritmu HITS (1)

$$H = \begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix} \qquad A = \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} \qquad M = \begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \end{array}$$



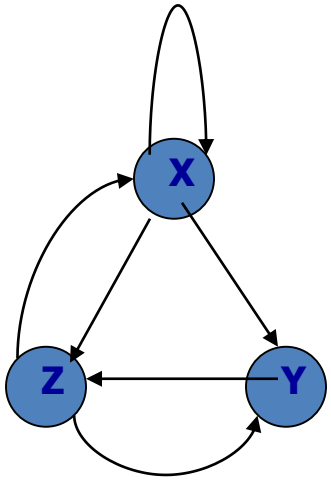
$$H_i = M * A_{i-1} \rightarrow H_i = M * M^T * H_{i-1}$$

$$A_i = M^T * H_{i-1} \rightarrow A_i = M^T * M * A_{i-1}$$

Príklad výpočtu algoritmu HITS (2)

$$M = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad M^T = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad M * M^T = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{bmatrix} \quad M^T * M = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

Iterácia 0 1 2 3 …



$$H = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 6 \\ 2 \\ 4 \end{bmatrix} \rightarrow \begin{bmatrix} 28 \\ 8 \\ 20 \end{bmatrix} \rightarrow \begin{bmatrix} 132 \\ 36 \\ 96 \end{bmatrix} = \begin{bmatrix} 0{,}79 \\ 0{,}21 \\ 0{,}57 \end{bmatrix}$$

X je najlepší hub

$$A = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 5 \\ 5 \\ 4 \end{bmatrix} \rightarrow \begin{bmatrix} 24 \\ 24 \\ 18 \end{bmatrix} \rightarrow \begin{bmatrix} 114 \\ 114 \\ 84 \end{bmatrix} = \begin{bmatrix} 0{,}63 \\ 0{,}63 \\ 0{,}46 \end{bmatrix}$$

X a Y sú najautoritatívnejšie stránky

Porovnanie PageRank a HITS

- **PageRank**

  (Google)

  – Vypočítava váhy pre všetky web stránky v databáze ešte pred zadaním dopytu

  – Vypočítava len autority

  – Triviálny a rýchly výpočet

- **HITS**

  (CLEVER)

  – Výpočet prebieha na množine stránok vrátených pre každý dopyt

  – Vypočítava autority a huby

  – Ľahký výpočet, ale ťažko realizovateľný v reálnom čase