Vyhľadávacie systémy využívajúce sémantické vyhľadávanie*

Martin Farkaš

Slovenská technická univerzita v Bratislave

Fakulta informatiky a informačných technológií

xfarkasm2@stuba.sk

30. september 2023

Abstrakt

Pri vyhľadávaní rôznych informácii na internete sa často stretávame so stránkami alebo článkami, ktoré by sme pri tvorbe našich projektov, prác alebo prezentácii nevyužili a to z jednoduchého dôvodu - ich obsah často nesúvisí s pojmom, ktorý práve hľadáme. Sem vstupuje pojem relevantnosť a konkrétne a jej uplatnenie v oblasti tvorby kvalitných vedeckých článkov alebo aj bakalárkych prác. Relevantnosťou vyhľadávaných informácii a článkov na internete sa práve zaoberá semantické vyhľadávanie. Je to pojem, ktorý je každému tvorcovi článkov, ale aj študentovi vysokej či strednej školy známy. O to zaujímavejšie sú systémy, ktoré využívajú sémantické vyhľadávanie, pretože každý z nich môže pracovať na inom princípe a výsledok niektorých z nich môže byť, čo sa relevantnosti týka, presnejší.

1 Úvod

Počas toho ako sa Internet postupne vyvíjal bolo čoraz ľahšie sa dostať k informáciám, ktoré sme potrebovali. V dnešnej dobe sme zahltení toľkým množstvom informácií, kde množstvo z nich je aj nepravdivých, že na vypracovanie nášho projektu by sme museli prehľadávať niekoľko desiatok stránok a zároveň kontrolovať či je daný zdroj overený. Samozrejme, na internete existuje množstvo stránok, ktoré nám poskytnú články z overených zdrojov, avšak ich obsah je často mimo tému/problém, ktorý riešime. Pri hľadaní relevantných článkov nám pomáhajú vyhľadávacie systémy, ktoré pracujú na princípe sémantického vyhľadávania.

Tento článok sa bude preto zaoberať:

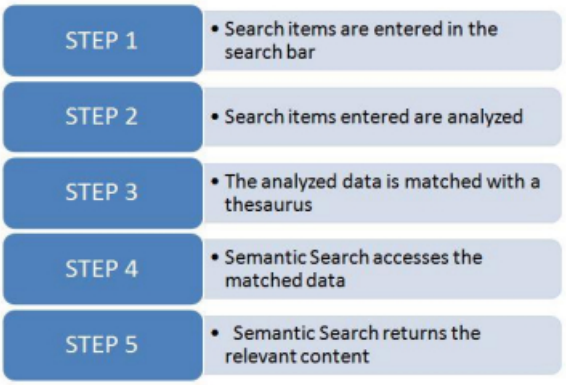
- Stručným vysvetlením pojmu sémantické vyhľadávanie 2
- Rozdiel medzi klasickým vyhľadávaním a sémantickým vyhľadávaním 3
- Konkrétnymi systémami a ich funkciami 4
- Systémom založeným na báze umelej inteligencie - "Semantic Scholar"

*Semestrálny projekt v predmete Metódy inžinierskej práce, ak. rok 2023/24, vedenie: Richard Marko

2 Pojem semantické vyhľadávanie

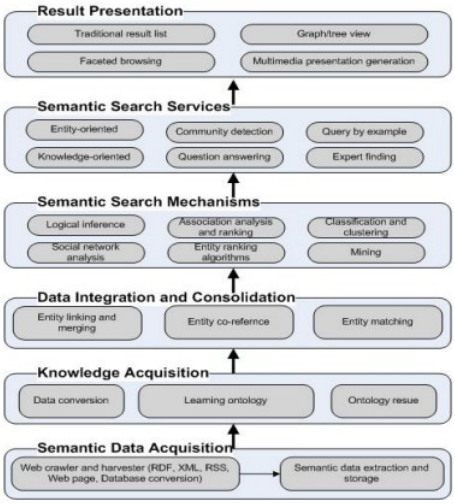
Aby sme mohli pokračovať v tomto článku, musia byť všetci čitatelia oboznámený s pojmom sémantické vyhľadávanie. Začnime pojmom sémantika."Sémantika sa sústreďuje najmä na skúmanie vzťahov medzi jazykovými výrazmi a predmetmi, na ktoré sa tieto výrazy vzťahujú a na tie vlastnosti a vzťahy výrazov, ktoré súvisia a ich vzťahmi k týmto predmetom.- [P.C01] S vynálezom prvých počítačov prišla aj myšlienka použiť sémantiku v počítačoch. Podľa [SR] sa za priekopníka v tejto oblasti považuje Robert W. Floyd, ktorý vo svojej práci opísal využitie sémantiky v počítačoch. Jeho práca obsahovala dizajn algoritmov použitých na vyhľadanie najefektívnejšej cesty v sieti alebo aj triedenie informácii. Z toho článok [SR] vyvodil zámer, že sémantické vyhľadávanie, na rozdiel od typických vyhľadávacích algoritmov, je založený na kontexte vyhľadávanej frázy, zámere aj jej podstate. Využíva viacero algorimov a metód, ako napríklad "keyword-to-concept mapping"čo môžeme preložiť ako vytváranie pojmových máp na základe kľúčových slov.

Princíp fungovania sémantického vyhľadávania je ukázaný na nasledovnom obrázku:



Obr. 1: Proces semantického vyhľadávania [SR]

Ako je na obrázku vidieť, proces vyhľadávania rozdeľuje autor článku [SR] na 5 častí. Prvým krokom je zadanie hľadaných slov. V ďalšom kroku sa tieto slova analyzujú a porovnávajú sa s "thesaurus"čo je kniha alebo lexikón so synonymami alebo slovami spolu súvisiacimi. V preposlednom kroku sa prístúpi k zhodným dátam a vyhľadávanie vyhodí vhodné výsledky v podobe stránok alebo článkov. Avšak článok [WW] rozdeľuje činnosť až na 6 častí ako vidieť na obrázku č. 2. Prvá časť poskytuje zbieranie neštruktúrovaných (internetové stránky), polo-štruktúrovaných (dáta v XML a databáze) a štruktúrovaných dát, avšak neštruktúrované a polo-štruktúrované dáta musia byť pretranformované na štruktúrované. Tu vstupuje do funkcie druhá časť, ktorá transformuje tieto dáta podľa určitých techník/metód na spracovanie dát ako napríklad konverzia dát. Tretia časť sumarizuje výsledky pre problémy ktoré sa mohli vyskytnúť pri 2. časti ako napríklad prípad kedy rôzne zdroje môžu poskytovať rôzne a často doplnkové informácie k istým problémom. Ďalšími časťami sú rôzne mechanizmy a služby semantického vyhľadávania a na koniec prezentácia výsledkov



Obr. 2: O trochu zložitejšia štruktúra vyhľadávania [WW]

3 Porovnanie s klasickým vyhľadávaním

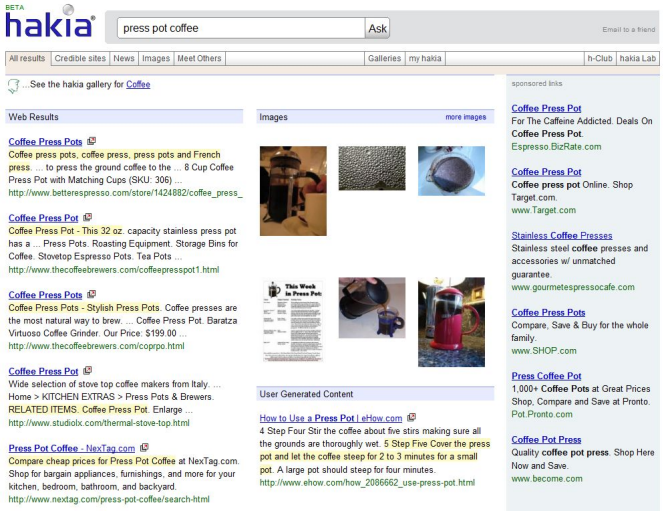
4 Konkrétne vyhľadávacie systémy

V tejto časti by som sa chcel povenovať konkrétnym vyhľadávacím systémom, ktoré pracujú na báze sémantického vyhľadávania, ich funkciám.

4.1 Hakia

Prvým systémom je, ako je aj z názvu vidieť, Hakia. Autor článku [JSS] uvádza, že tento vyhľadávací systém bol vyvynutý jadrovým vedcom Riza Berkan a ekonómom Kouri v roku 2004. Taktiež uvádza, že sa Hakia zameriava skôr na význam slov a nie priamo na slová a slovné frázy. Článok [SR] navyše uvádza, že tento systém vyzýval používateľa zadávať nielen slová, ale aj otázky, frázy alebo celé vety. Výsledky vyhľadávania boli rozdelené do kategórii Web, Správy, Blogy, Videá a mohli sme si ich usporiadať podľa relevantnosti a dátumu. Oba články [GS] a [SR] uvádzajú, že Hakia bola založená na troch technológiách. Prvou bola OntoSem, ktorú autori článkov opisujú ako lingvistickú databázu kde sú slová sú klasifikované do rôznych "významov"ktoré vyjadrujú. Myslím si však, že technológiu OntoSem opisuje článok [MMB] najlepšie a to ako prostredenie spracovávajúce text, ktoré za vstup berie absolútny text (z anglického slova ünrestricted text") a na výstupe vynáša von jeho morfologickú, syntaktickú a sémantickú analýzu. Druhá technológia s názvom QDEX (Query indexing technique), ako autori vyššie uvedených článkov píšu, mala za úlohu zozbierať všetky možné významy súvisiace s obsahom nášho vyhľadávania. Poslednou je algoritmus Semantic Rank, ktorý nezávisle zoraďoval výsledky vyhľadávania, čo používateľovi ušetril čas a ľahko identifikoval informácie z spoľahlivých stránok. Okrem toho mal systém Hakia aj plno iných funkcií ako uvádza napríklad autor článku [JSS], kde uvádza, že Hakia presentoval niekoľko návrhov pred tým ako sme stlačili tlačidlo "Vyhľadať", zvýraznil slová alebo vety, ktoré dávali odpo-

veď na nami hľadaný problém a taktiež je v danom článku uvedené, že Hakia používal aplikáciu "Meet Others", ktorá umožňovala používateľom stretnúť sa a diskutovať o dôležitých problémoch. Na druhej strane má Hakia aj pár limitácii ako uvádza autor článku [AK]. Medzi jeho nevýhody patrí napríklad fakt, že všetko neindexuje, potrebuje iné vyhľadávacie systémy na jeho funkciu.



Obr. 3: Náhľad na používateľské prostredie Hakia [Joh]

4.2 Kngine

Ďalším z radu systémom, ktoré využívajú sémantické vyhľadávanie patrí Kngine. Tento systém má toho veľa spoločného so sytémom Hakia ako uvádza autor článku [SR] ako napríklad na vstupe môžeme zadávať otázky. Článok taktiež uvádza, že výsledky sú rozdelené na výsledky vyhľadávania na webe a na výsledky vyhľadávania medzi obrázkami. Okrem toho používa ako uvádza aj autor vyššie uvedeného článku, schopnosť sám sa učiť. To znamená, že sa neustále učí a rozvíja. Aktuálne obsahuje viac ako 8 miliónov konceptov a to je miesto kde podľa autora článku [GS] je jeho najväčšia sila. Okrem toho článok [AK] uvádza, že je multijazyčný a okrem toho umožňuje používateľovi vyhľadávať paralelne.

4.3 Powerset

Ako autor článku [SR] uvádza, v roku 2005 vznikla firma, ktorá škonštruovala daný vyhľadávací systém s myšlienkou spraviť vyhľadávanie ľahšie a viac intuitívne. Činnosť systému Powerset sa podľa daného článku sústreďuje iba na jedinú vec a to na spracovanie natívneho/prirodzeného jazyka aby rozumel podstate otázky alebo vyhľadávanej fráze. Taktiež článok [GS] uvádza, že všetky výsledky vyhľadávania pochádzajú zo stránky Wikipedia.

4.4 Swoogle

Ďalším z tých populárnejších vyhľadávacích systémov je Swoogle. Autor článku [AK] píše, že Swoogle je založený na "crawlerovi". Swoogle využíva "crawlera"na objavovanie RDF a HTML dokumentov a následne extraktuje dáta a zhodnotí spojitosti medzi dokumentmi. Ak by sme nevedeli čo pojem "crawler znamená, článok [MAK] nám to pekne vysvetlí . Opisuje ho ako program alebo softvér, ktorý prehliada internet systematickým a automatickým spôsobom. Taktiež autor článku píše, že sú používané v podstate na vytváranie replík navštívených stránok, ktoré sú neskôr spracované vyhľadávacím systémom, ktorý zaindexuje stiahnuté stránky ktoré pomôžu v rýchlom vyhľadávaní. Článok [AK] ho opisuje aj ako vyhľadávací systém založený na obsahu, ktorý zvykol analyzovať, objavovať a indexovať vedomosti získané z internetu a okrem toho aj ako vyhľadávací systém založený na algoritme, ktorý hodnotí internetové stránky podľa rôznych kritérii a OWL jazyku. Taktiež autor daného článku píše o jeho limitáciách medzi ktoré patrí zlé indexovanie dokumenov a slabá doba odozvy pri hľadaní výrazov.

4.5 Sensebot

Jedným z posledných vyhľadávacích systémov je Sensebot. Ako autor článku [SR] uvádza, Sensebot používa ťaženie textu (z anglického spojenia "text mining") na analýzu internetových stránok a rozpoznáva dôležité sémantické koncepcie. Potom vykonáva niekoľko dokumentový prehľad obsahu aby následne vygeneroval precízny súhrn výsledkov vyhľadávania. Autor článku [KSG] opisuje priebeh hľadania výsledkov o trochu inak. Keď používateľ zadá výraz na hľadanie, Sensebot bude najprv hľadať niekoľko stránok, ktoré sa zhodujú s hľadaným výrazom. Avšak predtým ako vráti používateľovi výsledky, urobí analýzu výsledkov pomocou už spomínaného "text miningu", ktorý autor opisuje ako technológiu vyhľadávajúcu kľúčové pojmy zahrnuté na stránkach. Následne je vykonávaná niekoľko dokumentová sumarizácia na vytvorenie súhrnu tém súvisiacim s používateľovým vstupom, ktorý je vrátený na výstupe. Ako vidieť, autori sa v hlavných koncepciách zhodujú. Autor druhého článku vysvetľuje danú problmatiku o trochu podrobnejšie.

4.6 DuckDuckGo

Posledným systémom využívajúci sémantické vyhľadávanie je DuckDuckGo. Najväčšou črtou tohto vyhľadávacieho systému je ochrana súkromia a osobných údajov. Ako majú uvedené aj na ich internetovej stránke duckduckgo.com je to bezplatný webový prehliadač ktorý nesleduje ani našu históriu vyhľadávania, ani históriu prehliadania. Povedal by som, že to je jeden z najviac populárnych vyhľadávačoch po gigantoch ako Google, Yahoo alebo Bing. Okrem toho je plný iných funkcií. Autor článku [GS] uvádza funkciu, že ak hľadaný výraz alebo slovo má viac významov, DuckDuckgo mu dá možnosť zvoliť si daný význam slova, ktorý pôvodne hľadal.

5 Semantic Scholar

6 Ešte dôležitejšia časť

7 Záver

Literatúra

[AK] Dr. Jawahar Thakur Amita Kumari. Semantic web search engines : A comparative survey. https://ijsrcseit.com/PDF.php?pid=CSEIT195115&v=5&i=1&y=2019&m=January-February.

[GS] Prof. M. Surendra Prasad Babu G. Sudeepthi, G. Anuradha. A survey on semantic web search engine. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=de8b37623b6a1ea2c3d103d2cf4182856227aead.

[Joh] Nathania Johnson. Meet the new hakia. https://www.searchenginewatch.com/2008/10/06/meet-the-new-hakia/.

[JSS] Vishal Jain, Gagandeep Singh, and Mayank Singh. Comparative study of search engine and semantic search engine: A survey.

[KSG] Patricia Anthony Vooi Keong Boo Kim Soon Gan1, Kim On Chin. Dbpedia based meta search engine. https://tost.unise.org/pdfs/vol4/no3/4x3x232x251.pdf.

[MAK] Sanjeev Kumar Singh Md. Abu Kausar, V. S. Dhaka. Web crawler: A review. https://research.ijcaonline.org/volume63/number2/pxc3885125.pdf.

[MMB] Sergei Nirenburg Marjorie McShane and Stephen Beale. The description and processing of multiword expressions in ontosem. https://www.researchgate.net/profile/Sergei-Nirenburg/publication/254848860_The_Description_and_Processing_of_Multiword_Expressions_in_OntoSem/links/0deec53beeff877d00000000/The-Description-and-Processing-of-Multiword-Expressions-in-OntoSem.pdf.

[P.C01] P.Cmorej. *Úvod do logickej syntaxe a sémantiky.* IRIS, Bratislava, 2001.

[SR] Debabrata Barik Surajit Goon Subham Roy, Akshay Modak. An overview of semantic search engines. https://www.ijrrjournal.com/IJRR_Vol.6_Issue.10_Oct2019/IJRR0012.pdf.

[WW] Andrzej Bargiela Wang Wei, Payam M. Barnaghi. Search with meanings:an overview of semantic search systems. http://www.bargiela.com/papers/a35.pdf.