POVEZAVA REZULTATOV ISKANJA SPLETNEGA INTELIGENTNEGA AGENTA S PODATKI POMEMBNIMI ZA POSLOVNE ODLOČITVE

Dejan Lavbič
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Tržaška 25, 1000 Ljubljana
Dejan.Lavbic@campus.fri.uni-lj.si

Povzetek

Svetovni splet predstavlja velik in praviloma neizkoriščen vir informacij, ki ga lahko koristno uporabimo v sistemih za poslovno odločanje.[1] Članek se ukvarja s problematiko učinkovitega iskanja informacij na spletu in povezavo rezultatov iskanja s podatkovnimi skladišči za potrebe OLAP analiz. Podaja osnovne koncepte za razvoj informacijskega sistema iskanja informacij na svetovnem spletu in koristno uporabo le-teh. Ideja temelji na dejstvu, da se je potrebno spletnega iskanja lotiti tako na sintaktični kot na semantični način, z uporabo za splet posodobljenih algoritmov za iskanje.

Abstract

LINKING SEARCH RESULTS OF WEB INTELLIGENT AGENT WITH DATA RELEVANT TO BUSINESS DECISION MAKING

World-Wide Web represents plentiful, unexploited source of information, which could be useful in Decision Support Systems, therefore this article is engaged in effective searching of Web information and linking the results with Data Warehouses for the use of OLAP analysis. We are trying to induct some basic concepts to developing Information Systems for information retrieval on the Web and useful usage. The idea is that Web has to be searched both semantically and syntactically using web aware information retrieval algorithms.

1. UVOD

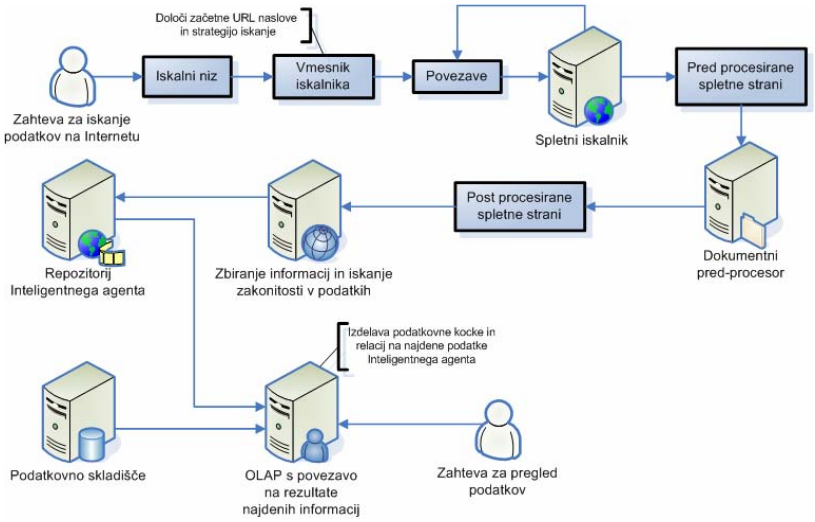
V zadnjem času smo priča zelo hitremu razvoju svetovnega spleta, zato želimo opozoriti na eksponentno rast objavljanja številnih in raznolikih vsebin na medmrežju. Slabost teh vsebin je predvsem način priprave vsebine, saj se le-ta zaradi lažjega vzdrževanja na spletu največkrat pretvarja iz statičnih strani in tekstovnih dokumentov v dinamično obliko na podlagi prednastavljenih šablon in vsebin, pridobljenih iz podatkovnih baz. V splošnem so na takšen način ustvarjene komercialne spletne strani, kot so iskalniki, elektronske trgovine, novičkarski portali itd.

Število spletni strani je tako pospešeno raslo, a te strani po drugi vsebujejo tudi veliko odvečnih in nepomembnih informacij. Primeri takšnih informacij so oglasi, menuji za pomoč pri brskanju po straneh, seznami storitev, izjave o pravicah za razmnoževanje in varnostni politiki, vsebina, označena s povezavami na razlage določenih pojmov itd. Pri sistematičnih spletnih straneh, kot so npr. strani z ažurnimi novicami, je za uporabnike prisotnost redundantnih informacij dobrodošla, saj jim omogoča navigacijo in iskanje določene vsebine z manj kliki in bližnjicami. Po drugi strani pa te odvečne povezave otežijo delo spletnim iskalnikom pri iskanju uporabnih informacij, ker so le-ti ponavadi pripravljeni tako, da indeksirajo in sprocesirajo vse, vključno z odvečno in nepomembno vsebino. V splošnem

[1] Decision Support System (DSS).

odvečne informacije na spletnih straneh niso tesno povezane z vsebino, kar posledično pripelje do problema nepravilne klasifikacije vsebine strani.

Ugotovimo lahko, da svetovni splet vsebuje zelo veliko uporabnih informacij, tudi takšnih za zmanjšanje entropije pri poslovnih odločitvah, obstaja le težava, kako se do njih dokopati in jih na prijazen in intuitiven način prikazati uporabniku – analitiku, ki jih lahko uporabi pri procesu odločanja.



Slika 1 : Koncept povezovanja

Namen članka je podati konceptualen opis možne rešitve iskanja informacij na spletu in rezultate učinkovito povezati z obstoječim sistemom odločanja v poslovanem sistemu. *Slika 1* prikazuje osnovni koncept povezovanja najdenih informacij na spletu s podatki, potrebnimi za poslovne odločitve. Koncept je zaradi morebitnih prilagoditev različnim ciljnim sistemom podan zelo modularno.

2. INTELIGENTNI AGENT

Inteligentni agent je program, ki išče (po medmrežju) in zbira vire iz svetovnega spleta. Uporablja se predvsem za zbiranje strani, indeksiranje pri spletnih iskalnikih, lahko pa se uporabi tudi za zbiranje informacij pri spletnem iskanju zakonitosti v podatkih.[2]

Delovanje inteligentnega agenta lahko opišemo kot vzdrževanje čakalne vrste URL naslovov, ki jih iskalnik namerava obiskati. Med procesom iskanja se trenutni URL naslov izloči iz čakalne vrste, obišče se stran, na katero povezava kaže, v čakalno vrsto pa se dodajo vse ali določena podmnožica izhodnih povezav, ki se nahajajo na trenutni strani. Zaradi performančnih razlogov inteligentni agenti ponavadi uporabljajo asinhroni V/I, kar jim omogoča obiskovanje več strani hkrati, ali pa so implementirani kot večnitni sistemi, kjer vsaka nit opravlja osnovni proces iskanja, vzporedno z drugimi.

[2] Web Data Mining.

3. DELOVANJE AGENTA

Inteligentni agent začne iskanje z množico preddefiniranih URL naslovov. Če poenostavimo: inteligentni agent najprej postavi to množico URL-jev v čakalno vrsto in jih označi kot prioritetne. Med delovanjem iz te vrste po nekem vrstnem redu pobira URL naslove in išče podatke med vsebino spletnih strani, išče nove URL naslove na obdelanih straneh in jih dodaja v čakalno vrsto. Ta proces se ponavlja, vse dokler se inteligentni agent ne odloči, da s svojim delom konča. Ob upoštevanju ogromne razsežnosti svetovnega spleta in glede na intenziteto spreminjanja vsebine le-tega, se ob tem postavljajo mnoga vprašanja:

- **Katere spletne strani naj inteligentni agent obišče?**

  V večini primerov inteligentni agent ne more obiskati vseh strani na spletu, saj celo največji iskalniki (Google, Yahoo ...) pokrivajo le majhen del celotnega spleta. Na podlagi tega spoznanja je zelo pomembno, da inteligentni agent najprej izbira »pomembne« strani z uporabo prioritet na nivoju URL naslovov v čakalni vrsti.

- **Kako naj inteligentni agent osvežuje strani?**

  Ko inteligentni agent obišče določeno število strani, je potrebno razmišljati o ponovnem obisku zaradi sprememb, ki so nastale od zadnjega obiska. Ker se vsebina spletnih strani spreminja zelo različno, se mora inteligentni agent odločiti, katero stran ponovno obiskati in katero izpustiti v določenem ciklu iskanja.

- **Kako zmanjšati obremenjenost na spletnih strežnikih?**

  Pri obiskovanju spletnih strani inteligentni agent zahteva na gostujočem sistemu sistemska sredstva, saj mora – npr. pri zahtevi po neki strani – gostujoči sistem le-to prebrati iz datotečnega sistema in jo prikazati, kar pa zahteva sistemska sredstva, zato mora inteligentni agent čimbolj zmanjšati svoj vpliv na delovanje gostujočih sistemov, pri čemer je priporočljivo, da upošteva Robots Exclusion Protocol.[3]

- **Kako proces iskanja paralelizirati?**

  Zaradi velikega obsega spleta je potrebno pomisliti na paralelizacijo iskanja predvsem v primeru potrebe po velikem številu strani v razumljivem časovnem okviru. Jasno je, da je potrebno poskrbeti za koordinacijo vseh iskalnikov s centralne točke, tako da različni agenti ne obiščejo istih strani večkrat.

3.1 Izbira strani

Za inteligentnega agenta je pomembno, da najprej obišče pomembnejše strani. Pri tem se moramo predvsem zavedati pomena »pomembnosti«, tj. kako agent deluje in kako najde dobre strani za obiskovanje. Pri procesu izbire strani je možnih več tehnik, vendar je ena najbolj uporabnih **metrika pomembnosti**, kjer poznamo več variacij:

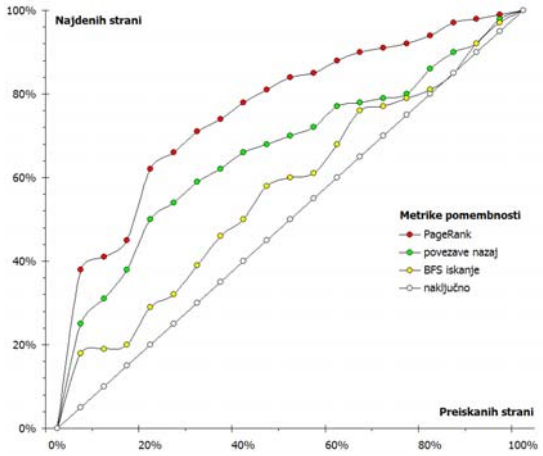
- Glede na **interes** določene skupine uporabnikov uporabljamo tekstovne podobnosti določenega dokumenta z referenčnim. V tem primeru moramo pregledati vsebino obeh dokumentov na podlagi n-dimenzionalnega vektorja $\langle w_1, w_2, \ldots, w_n \rangle$, kjer se pod w_i nahaja

[3] Skrbnik spletnega mesta lahko s pripravo ustrezne datoteke http://.../robots.txt določi katere strani spletni iskalnik pri iskanju ne obišče.

beseda iz slovarja in če se le-ta v dokumentu nahaja, dobi i-ta lokacija vrednost 1, sicer pa 0. Nato med obema vektorjema izvedemo skalarni produkt in dobimo oceno.

- Glede na **priljubljenost**, kjer ugotavljamo, kako popularna je stran. Preprost način je štetje povezav nazaj.[4] Le-te sestavljajo množico povezav, ki se nahajajo na poljubnih drugih straneh (različnih od dane strani) in kažejo na dano stran.

- Glede na **lokacijo**, kjer se ne uporablja odvisnost od vsebine, ampak funkcija odvisnosti od lokacije. Kot primer lahko navedemo .com URL naslove, ki so v splošnem nekoliko bolj uporabni in zato tudi dobijo višjo oceno. Prav tako uporabno je tudi število znakov /, ki se pojavljajo v URL naslovu, kjer večje število pomeni slabšo oceno.



Slika 2 : Uporaba različnih metrik pomembnosti pri postopku iskanju po spletu

Slika 2 [3] prikazuje uporabo treh različnih matrik pomembnosti pri izbiri naslednje strani za iskanje, kjer je jasno razvidno, kako lahko izbira primerne matrike pomembnosti, izdatno pohitri postopek iskanja.

4. REPOZITORIJ INTELIGENTNEGA AGENTA

Pod izrazom repozitorij inteligentnega agenta si predstavljamo prilagodljiv sistem za shrambo in urejanje velikega števila spletnih strani. V grobem mora repozitorij opravljati le dve osnovni funkciji:

- vmesnik za shranjevanje spletnih strani, ki jih agent najde

- vmesnik za dostop do shranjenih spletnih strani

Ker repozitorij skrbi za zelo veliko zbirko »podatkovnih objektov« – spletnih strani, je konceptualno zelo podoben ostalim sistemom za shranjevanje in urejanje podatkov (datotečni sistemi, podatkovne baze ...), vendar ne potrebuje implementiranih nekaterih funkcionalnosti (transakcije, pisanje dnevnika, direktorijska struktura ...), ki jih ti sistemi morajo zagotavljati, tako da se lahko bolj osredotočimo na vidike hitrejšega dostopa in skalabilnosti.

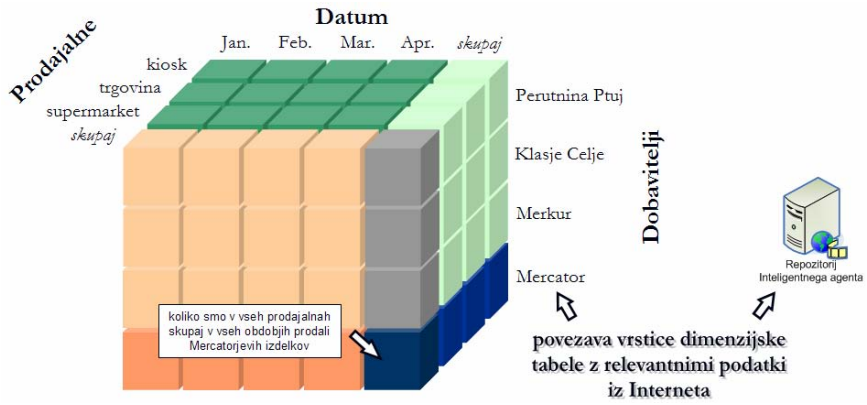
---

[4] Povezave, ki kažejo na dano stran.

5. PODATKOVNO SKLADIŠČE, OLAP IN POVEZAVA Z ZUNANJIMI VIRI PODATKOV

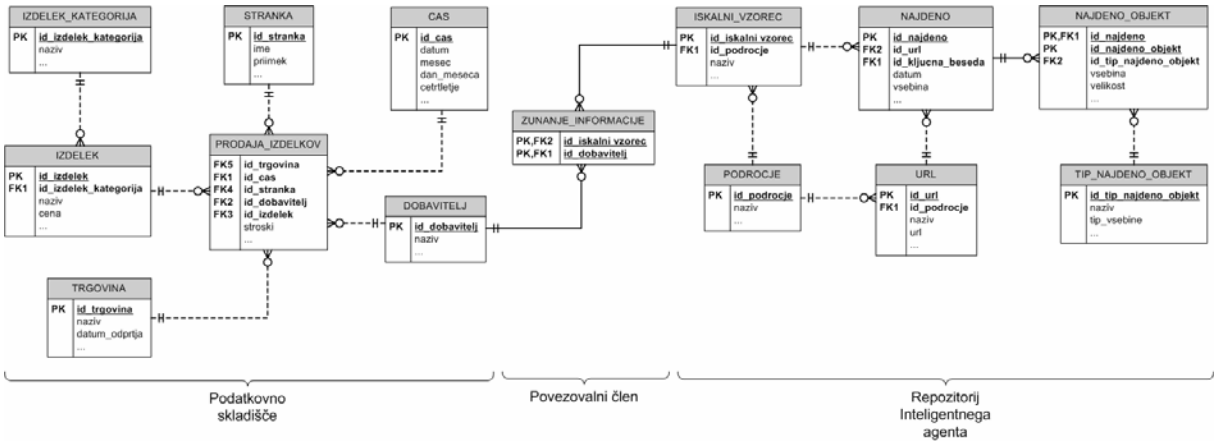
Pod pojmom podatkovno skladišče razumemo kopijo transakcijskih podatkov, posebno strukturiranih za poganjanje poizvedb in analiz [7] oz. analitično podatkovno bazo, ki je temelj odločitvenim sistemom.

Uporabnik podatkovnega skladišča je nosilec odločanja, ki ima predvsem poslovno znanje in mu je tehnično poznavanje problema sekundarnega pomena. Prvotna naloga takšnega analitika je definiranje in iskanje informacij, povezanih z odločitvenim mehanizmom združbe, zato je pomembno gledati na uporabnost podatkovnega skladišča z vidika analitika, ki za možne poslovne odločitve potrebuje čim več informacij iz tekočega poslovanja (podatkovne kocke, poročila ...) in tudi informacije, vezane na poslovno odločitev iz drugih virov, kot je npr. svetovni splet, dnevno časopisje itd.



Slika 3: Primer podatkovne kocke in povezave podatkov dimenzijske tabele s podatki v repozitoriju inteligentnega agenta

Slika 3 podaja primer iz prakse v zvezi s prodajo izdelkov (spremljanje dobička) določenih dobaviteljev po različnih tipih prodajaln v različnih časovnih obdobjih. Če se npr. odločamo, ali bomo izdelke od nekega dobavitelja kupovali še v prihodnje ali pa je čas za spremembe, si lahko odločitev olajšamo z zmanjšanjem negotovosti tako, da iz repozitorija inteligentnega agenta pregledamo podatke, ki so vezani na določenega poslovnega partnerja in iščemo morebitne razloge za nenadno spremembo njegovih cen.



Slika 4: Poenostavljen podatkovni model za povezavo repozitorija inteligentnega agenta in podatkovnega skladišča

Eden izmed možnih načinov implementacije povezave podatkovnega skladišča z repozitorijem inteligentnega agenta je podan na s*liki 4*, kjer se v levem delu diagrama nahaja shema podatkovnega skladišča s tabelo dejstev in pripadajočimi dimenzijskimi tabelami. Diagram v desnem delu pa prikazuje podatkovni model repozitorija inteligentnega agenta, kjer so shranjene vse pomembne in s problemsko domeno povezane spletne strani oz. vsebina le-teh in vsi pripadajoči objekti (grafični elementi, oblikovni vzorci ...). Entiteta ZUNANJE_INFORMACIJE predstavlja povezovalni člen zgoraj omenjenih komponent, ki konceptualno poveže določeno dimenzijo (v našem primeru smo se odločili za DOBAVITELJ, sicer pa je v splošnem izbira poljubna) z iskalnimi vzorci, na podlagi katerih inteligentni agent izvaja iskanje po svetovnem spletu.

6. ZAKLJUČEK

Dandanes je osnova mnogih informacijsko povezanih opravil uspešno preiskovanje svetovnega spleta. Uspeh spletnega iskalnika je odvisen od učinkovitega izločevanja informacij iz množice spletnih strani. S*lika 1* prikazuje koncept povezovanja najdenih informacij na spletu s podatki, potrebnimi za poslovne odločitve. Pri doseganju optimalnih rezultatov iskanja je kritična predvsem komponenta za zbiranje informacij in zakonitosti v podatkih iz s*like 1*. Z ustreznimi algoritmi je potrebno določiti dele strani, kjer se iskana informacija dejansko nahaja, ter odstraniti redundantne in nepomembne podatke.

7. VIRI IN LITERATURA

[1] HUNG-YU, Kao, SHIAN-HUA, Lin, JAN-MING, Ho, MING-SYAN, Chen, Entropy-Based Link Analysis for Mining Web Informative Structures, Proceedings of the eleventh international conference on Information and knowledge management, November 2002.

[2] CHUNG, Chiasen, L. A. CLARKE, Charles, Topic-Oriented Collaborative Crawling, Proceedings of the eleventh international conference on Information and knowledge management, November 2002.

[3] ARASU, Arvind, CHO, Junghoo, GARCIA-MOLINA, Hector, PAEPCKE, Andreas, RAGHAVAN, Sriram, Searching the Web, ACM Transactions on Internet Technology (TOIT), Stanford University, Avgust 2001.

[4] CESARANO, Carmine, D'ACIERNO, Antonio, PICARIELLO, Antonio, An Intelligent Search Agent System for Semantic Information Retrieval on the Internet, Proceedings of the fifth ACM international workshop on Web information and data management, November 2003.

[5] MILLER, Rob, WebSPHINX: A Personal, Customizable Web Crawler, http://www-2.cs.cmu.edu/~rcm/websphinx/.

[6] INMON, W. H., Building the Data Warehouse, Third edition, Wiley Computer Publishing, 2002.

[7] KIMBALL, Ralph, The Data Warehouse toolkit, Second edition, Prentice Hall, 1996.