ISSN 1392-8619 print/ISSN 1822-3613 online
ŪKIO TECHNOLOGINIS IR EKONOMINIS VYSTYMAS
TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY
http://www.tede.vgtu.lt
2007, Vol XIII, No 3, 237–243

PROGRAMINIO AGENTO, NAUDOJANČIO NEKILNOJAMOJO TURTO ONTOLOGIJĄ, TAIKYMAS KAIP DUOMENŲ ŠALTINIS SPRENDIMŲ PARAMOS SISTEMAI

Darius Jurkevičius, Algirdas Laukaitis

Vilniaus Gedimino technikos universitetas, Fundamentinių mokslų fakultetas,
Saulėtekio al. 11, LT-10223 Vilnius, Lietuva
El. paštas fmf@fm.vgtu.lt

Įteikta 2007-02-27; priimta 2007-09-10

Santrauka. Siekiant didesnio darbo našumo žinių visuomenėje, greitas ir kokybiškas informacijos gavimas tampa svarbiu faktoriumi, kuris lemia organizacijos sėkmę atkaklioje konkurencinėje kovoje. Norint pasiekti užsibrėžtų tikslų, šiuo metu jau nebeužtenka naudotis informacijos paieškos internete sistemomis, tokiomis kaip *Google* ar *Yahoo*. Šiame straipsnyje pateikiama programinio agento architektūra ir jo realizacijos metu atlikti tyrimai, kurie parodo, kad siūlomi sprendimai leidžia surinkti ir sutvarkyti nekilnojamojo turto skelbimus daug efektyviau, nei tai darytų žmogus, naudojantis tik interneto informacijos paieškos sistemas. Nekilnojamojo turto rinka pasirinkta neatsitiktinai. Tai viena iš dinamiškiausių rinkų Lietuvoje, kurioje greitas informacijos pateikimas lemia sėkmingus sprendimus. Nors pateikti sprendimai ir orientuoti į nekilnojamojo turto segmentą, tačiau daugelis koncepcijų gali būti sėkmingai pritaikyti ir kitiems ekonomikos segmentams. Pagrindinis akcentas pateiktoje programinio agento architektūroje yra ontologijos naudojimas surenkant ir sutvarkant nekilnojamojo turto informaciją. Be programinio agento konceptualių sprendimų, straipsnyje pateikiami ir konkretūs realizacijos moduliai, kurie buvo sukurti šio projekto metu.

Reikšminiai žodžiai: nekilnojamasis turtas, ontologija, informacijos paieška, informacijos išgavimas, programinis agentas.

APPLICATION OF THE PROGRAM AGENT, THAT USES REALTY ONTOLOGY, AS THE DATA SOURCE FOR THE DECISION SUPPORT SYSTEM

Darius Jurkevičius, Algirdas Laukaitis

Vilnius Gediminas Technical University, Faculty of Fundamental Sciences,
Saulėtekio al. 11, LT-10223 Vilnius, Lithuania
E-mail: fmf@fm.vgtu.lt

Received 27 Februaury 2007; accepted 10 September 2007

Abstract. For a better operation efficiency of the knowledge society, fast and qualitative information reception is an important factor that determines the organisational success in a competitive contest. It is not enough to use the Internet search systems such as Google or Yahoo in reaching the established goals. In this article we present the architecture of the program agent and the results of his implementation studies that show how the suggested decisions allow picking and arranging realty advertisements more effectively than those people who apply only Internet information searching systems. The realty was chosen because it is one of the most dynamic spheres in Lithuania, when successful decisions are determined by a fast information. Though the suggestions are oriented to the realty section, different conceptions could be adapted to other economy segments. The main accent of the shown programme agent architecture is using the ontology picking and arranging the realty information. In this article we present not only conceptual solutions of the programme agent, but also specific realisation modules developed during this project.

Keywords: realty, ontology, search information, information retrieval, programme agent.

1. Įvadas

Filosofinius pagrindus, kuriais remiasi dabartinis žinių ir informacijos valdymas Vakarų civilizacijoje padėjo graikų filosofai Sokratas, Platonas ir Aristotelis. Aristotelio kategorijos ir dabar žinių valdymo ekspertų cituojamos kaip išeities taškas, nuo kurio mes pradedame klasifikuoti visus mus supančius objektus. Ilgą laiką ontologijos buvo tik filosofijos tyrinėjimo objektas tačiau šiame interneto amžiuje, kai informacijos kiekis, kuris gali pasiekti mus per interneto technologijas, auga eksponentiniu greičiu ir kai atsirinkti reikiamą informaciją tampa vis sunkiau, ontologijos tampa visų pirma informatikos specialistų tyrinėjimo objektu. To patvirtinimu gali būti OWL – senatinio interneto ontologijų aprašymo kalba, leidžianti „susišnekėti" programiniams agentams semantiniu lygmeniu. Šiame straipsnyje mes pateikiame sprendimus, kaip gauti tikslesnius paieškos rezultatus integruojant nekilnojamojo turto ontologiją į informacijos paieškos sistemas.

Labai dažnai žmonės, norintys greitai surasti nekilnojamąjį turtą ir dėl tam tikrų priežasčių nenorintys kreiptis pagalbos į nekilnojamojo turto agentūras, paieškos šaltiniu dažniausiai pasirenka internetą. Bet neretai atsitinka taip, kad norint sutaupyti laiką, būna priešingai. Bandymai pasirinkti geriausią sprendimą lieka neįgyvendinti, nes žmogus negali vienu metu įvertinti tokį kiekį informacijos. Ypač tai aktualu vertinant nekilnojamąjį turtą, kai reikia atsižvelgti į daugelį kriterijų. Daugelis pastaruoju metu atliktų vartotojų elgsenos ir sprendimų priėmimo tyrimų rodo, kad žmogus „*tingi*" pagrįsti savo sprendimus daugeliu kriterijų ir priima sprendimus, remdamasis vienu ar dviem kriterijais ir dėl to lieka nepatenkintas savo sprendimu. Taigi kyla akivaizdus poreikis, leidžiantis automatizuoti sprendimų priėmimo mechanizmą ir lemiantis daugeliu kriterijų.

Keičiantis vartojimo visuomenės įpročiams, šiuo metu internete galima vis dažniau rasti informacijos, kurios nėra kituose informacijos pateikimo subjektuose, kaip antai spauda, radijas ar televizija. Geras pavyzdys yra skelbimų laikraštis „Alioreklama", kurio turinys atkartojamas interneto svetainėje *www.alioreklama.lt.* Skelbimą pateikti į šį laikraštį galima dviem būdais: paskambinus į laikraščio redakciją arba įvedus skelbimo turinį laikraščio svetainėje. Tačiau čia pateikiami skelbimai – tik lašas interneto informacijos jūroje.

Deja, internete egzistuoja daug svetainių, kurių pateikiama medžiaga dažnai keičiama ir informacija dažnai turi nekonstruktyvų turinį. Pavyzdžiui, skelbimų svetainės, kuriose pateikiami skelbimai apie nekilnojamąjį turtą. Jei panaudotume paieškos programą arba paieškos agentą ir surinktume šiuos skelbimus, negautume naudingos informacijos nekilnojamojo turto analizei, nes tai būtų tik duomenų rinkinys. Šis metodas tiktų skelbimams surinkti, bet jis netinka jiems suprasti ir panaudoti analizei. Tam, kad šie duomenys būtų naudingi, iš jų reikia išrinkti tam tikrus kriterijus, naudojamus nekilnojamojo turto analizei. Tai yra todėl, kad nėra bendro standarto duomenims atvaizduoti, ir naudojamas HTML kalbos tvarkinys.

Taigi mes išskiriame kitus interneto kaip informacijos šaltinio naudojimo privalumus ir trūkumus.

Pagrindiniai privalumai galėtų būti šie:

- informacija, pateikta internete, galima lengvai manipuliuoti, kadangi ji yra pateikta elektroniniu formatu;
- turint asmeninį kompiuterį ir interneto ryšį, informacija yra lengvai pasiekiama, nereikia įsigyti laikraščių ar kitų informacijos šaltinių;
- patogi pateikimo forma;
- patalpintas skelbimas atsiranda internete nedelsiant, ir jį galima iškart peržiūrėti.

Kitus faktorius mes įvardijame kaip trūkumus:

- informacija pateikta skirtinguose puslapiuose;
- informacija pateikta natūralia kalba (yra labai daug klaidų);
- informacija dažnai kartojasi;
- informacija gali būti neišsami.

2. Problemos aprašymas

Spręsdami informacijos paieškos uždavinį, išskiriame šias problemas:

- informacija internete išbarstyta skirtinguose puslapiuose. Išskiriamos tokios jų kategorijos, kaip elektroninės skelbimo lentos; nekilnojamojo turto agentūrų svetainės. Nėra tikslinga vykdyti paiešką visame internete. Pasinaudojus nuorodomis iš katalogų, kur jie surūšiuoti pagal veiklą, galima rasti daug svetainių, kuriose pateikiama informacija apie nekilnojamąjį turtą. Pagrindinės svetainės, kuriose yra skelbiama reikiama informacija pagal *Google* svetainės duomenis, yra *http://www.aruodas.lt*, *http://www.skelbiu.lt*, *http://www.domoplius.lt*, *http://www.alioreklama.lt*, *http://www.skelbimai.lt*, *http://www.edomus.lt*, *http://www.enamai.lt*, *http://www.city24.lt*, *http://www.namai.lt*; *http://www.muge.lt*;
- informacija dažnai kartojasi (apie 30–50 % skelbimų svetainėse yra pasikartojanti informacija. Šis skaičius priklauso nuo interneto svetainės populiarumo, populiariausiose svetainėse informacija kartojasi dažniau). Šiam trūkumui pašalinti turi būti atlikta pasikartojančių skelbimų šalinimo procedūra;
- informacija kartojasi ir kai pasikeičia kuris nors skelbiamo nekilnojamojo turto parametras, dažniausiai keičiama kaina;
- nekilnojamojo turto agentas turi dirbti automatiniu režimu, ieškoti ir rinkti informaciją iš skelbimų svetainių, pvz., kas parą (šis laikas nustatomas eksperimentiniu būdu);
- siekiame, kad programinis informacijos paieškos agentas ne tik surinktų informaciją iš tam tikrą struktūrą turinčių skelbimų, bet kad ir galėtų rasti ir identifikuoti norimą informaciją iš teksto, kuriame nekilnojamojo turto informacija skelbiama tik kaip šalutinė informacija, o pats tekstas ir jo turinys orientuotas į kitą tematiką.

Siekdami išspręsti iškeltas problemas toliau pateiksime siūlomos sistemos architektūrą.

3. Sistemos architektūra

3.1. Nekilnojamojo turto agento architektūra

1 pav. pateiktoje architektūroje yra parodyti šie nekilnojamojo turto informacijos paieškos programinio agento moduliai:

paieškos agento funkcijas atliekantis *Web ScraperPlus+* programinis įrankis;

kriterijų išgavimo agentas;

pasikartojančių skelbimų šalinimo agentas;

agentas, kuris skelbimus perkelia iš DB į katalogą;

agentų veiksmus koordinuojantis agentas.

Architektūros pranašumai:

- išgaunami skelbimai, susiję tik su nekilnojamuoju turtu iš žinomų šaltinių. Tai neapkrauna tinklo;
- dažniausiai pasikartojantiems skelbimams šalinti naudojami duomenų bazės trigeriai;
- kriterijų išgavimo ir pasikartojančių (pagal prasmę, pavyzdžiui, pasikeitus NT kainai, paliekamas naujausias įrašas) skelbimų šalinimo procesuose naudojama ta pati nekilnojamojo turto ontologija;
- nurodyti svetainės schemą ir duomenų išgavimo taisykles reikia tik vieną kartą, nes interneto svetainių schemos dažnai nesikeičia.

Trūkumai:

- reikia nurodyti svetainės schemą ir duomenų išgavimo taisykles.

3.2. Duomenų išgavimas iš interneto svetainių

Duomenims išgauti iš interneto svetainių yra naudojamas programinis įrankis *Web Scraper Plus+*, kurio pagrindiniai mokuliai pavaizduoti 2 pav.

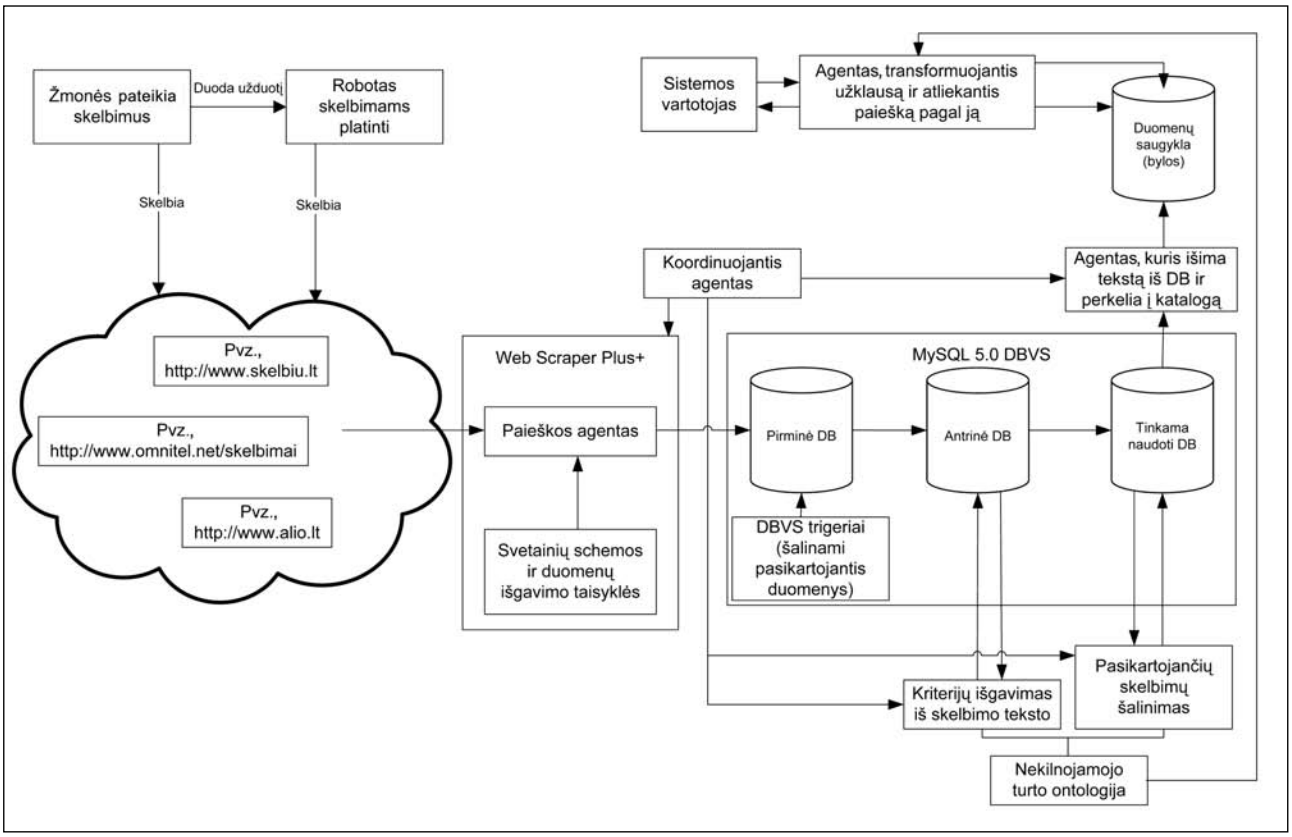
Web Scraper Plus+ programinio įrankio pasirinkimą lėmė šie privalumai:

- duomenų gavyba pagal šabloną;
- *web crawler*;
- automatinis formos pateikimas;
- duomenų kokybės sekimas;
- duomenų srauto valdymas;
- duomenų transformacija;
- galimybė informacijos išgavimo procesą leisti foniniu režimu.

Išgautiems duomenims saugoti yra parinkta *MySQL 5.0* duomenų bazės valdymo sistema, nes šioje versijoje yra realizuotas trigerių palaikymas. Trigeriais įgyvendinta pasikartojančių įrašų šalinimo funkcija. Šis sprendimas leidžia atsikratyti duomenų, kuriuos pateikia robotai skelbimams platinti. Paliekama tik naujausia informacija.

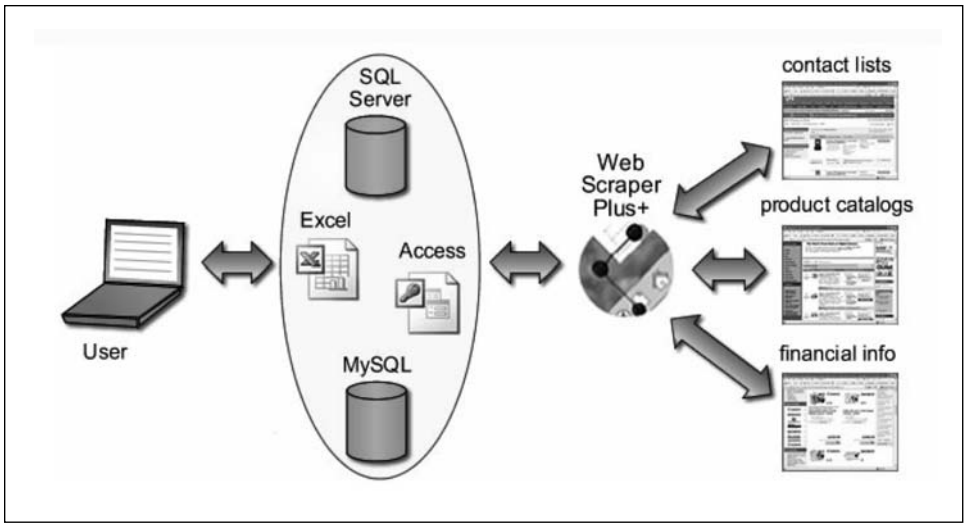
3.3. Nekilnojamojo turto ontologija

Tam, kad paieškos procese būtų galima atskirti pasikartojančius skelbimus ir iš jų išrinkti reikiamus kriterijus, naudojame OWL standartu aprašytą nekilnojamojo turto ontologiją. Web ontologijų kalba OWL [1, 2, 3] – tai kalba, nu-



1 pav. Nekilnojamojo turto agento architektūra, naudojanti ontologiją

Fig 1. The architecture agent of realty which uses ontology



2 pav. Duomenų gavybos procesas naudojant *Web Scraper Plus+*

Fig 2. Data retrieval process with Web Scraper Plus+

statanti ir parodanti WEB ontologijas. Formali semantika OWL aprašo, kaip išgauti logines išvadas, turint tokią ontologiją, t. y. gauti faktus, kurie neaprašyti ontologijoje, bet egzistuoja jos semantikoje.

Kodėl pasirinkome OWL? OWL WEB ontologijų kalba teikia didesnį žodyną savybėms ir klasėms aprašyti: aprašo ryšius tarp klasių (pavyzdžiui, neturintys susikirtimų), kardinalumą (pavyzdžiui, „tik vienas“), lygybes, išsamesnį savybių aprašymą, savybių charakteristikas (pavyzdžiui, simetriją) ir nepaminėtas klases. OWL ontologijų kalba buvo sukurta naudoti programose, kurios turi dirbti su informacijos turiniu, o ne teikti informaciją vartotojui. OWL pagerina galimybę automatiškai interpretuoti interneto turinį, lyginant su tuo, ką gali XML, RDF ir RDF schema. Tai užtikrinama tuo, kad OWL teikia papildomas informacijos aprašymo galimybes kartu su formalia semantika. Dar viena priežastis pasirinkti OWL yra platus ontologijų kūrimo nemokamų programinių įrankių pasirinkimas.

Nekilnojamojo turto ontologijai sukurti yra parinktas nemokamas *Protege 3.2 beta* programinis įrankis (3 pav.). *Protege 3.2 beta* gali dirbti su skirtingais ontologijų įrašymo formatais, tokiais kaip *XML*, *OWL/RDF database*, *OWL/RDF files, RDF files*.

3.4. Ontologijos naudojimas kriterijams gauti iš skelbimų teksto

Naudojant GATE [4, 5] programinį produktą buvo analizuojamas tekstas. Ši analizė buvo reikalinga ontologijai sukurti, sudarant terminų sąrašą.

Kriterijų išgavimo agento vykdomi veiksmai, kai yra išgaunami kriterijai yra šie (4 pav.) [6–11]:

- skelbimo nuskaitymas iš antrinės duomenų bazės kortežo;
- ontologijų paieška skelbimo tekste naudojant nekilnojamojo turto ontologiją;
- surastų ontologijų įterpimas į kortežo stulpelius. Kiekvienas stulpelis atitinka tam tikrą ontologiją, pavyzdžiui, ontologija plotas – jos atitikmuo 120 kv. m (atliekamas veiksmas: įrašyti 120 į stulpelį „plotas“);
- suradę visus galimus kriterijus ir juos įrašę į kortežą, pereiname prie kito skelbimo.

3.5. Nekilnojamojo turto ontologijos naudojimas apdorojant vartotojų užklausas

Sistemos vartotojai suformuluoja užklausas natūralia kalba.

Pavyzdžiui, yra užrašoma tokia užklausa: *Perku 3 kambarių butą Vilniuje*. Ši užklausa perduodama agentui, kuris ją transformuoja naudodamas nekilnojamojo turto ontologiją į SQL užklausą (5 pav.). Analizuojant užklausą yra praleidžiami visi žodžiai, kurie neturi atitikmens nekilnojamojo turto ontologijoje. Suformuota SQL užklausa yra įvykdoma ir rezultatai pateikiami vartotojui.

4. Tyrimo eiga

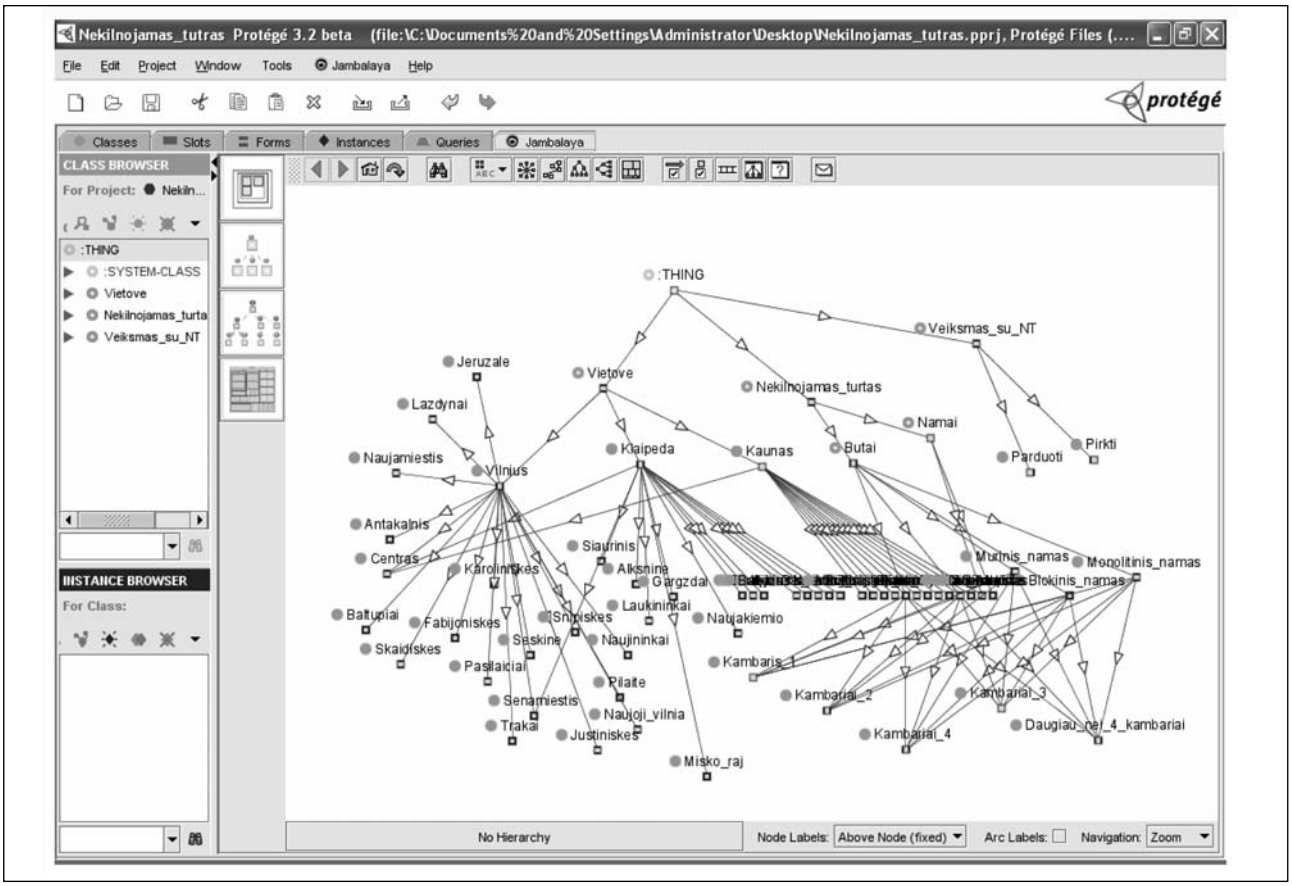
Pradėjus realizuoti prototipą buvo pastebėtos tokios problemos:

- kiek atsiranda naujos informacijos per laiko tarpą (pvz., per dieną). Tai reikia žinoti, norint neimti jau turimos informacijos;
- pašalinti pasikartojančias nuorodas.

Pirma problema išsisprendė nustačius didelį nuorodų į skelbimus skaičių (2000 vnt.). Stebint rezultatus per vieną savaitę buvo nustatytas maksimalus naujų nuorodų skaičius (356 vnt.). Taigi paieškos agentas buvo užprogramuotas išgauti maksimaliai 400 nuorodų į skelbimus per dieną.

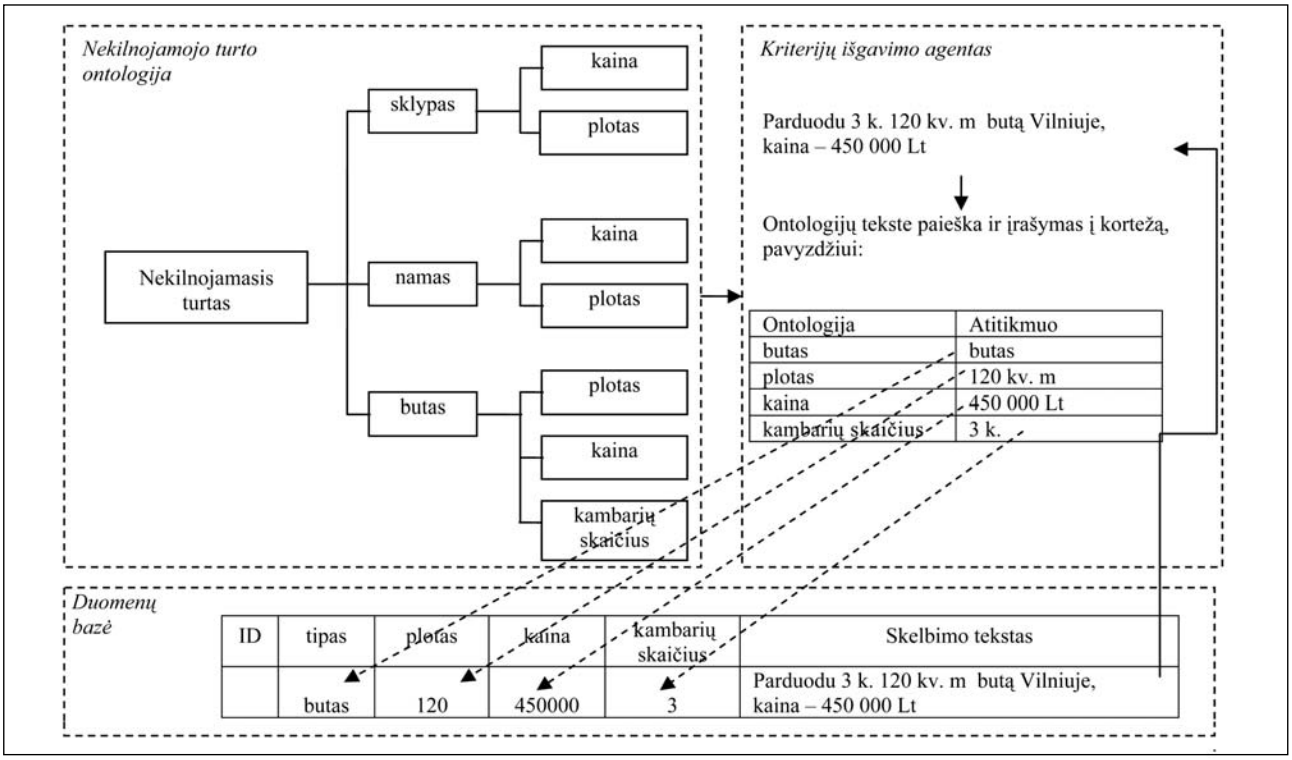
Norint iš tų 400 nuorodų palikti tik naująsias, visos buvo lyginamos su jau turimomis, sutapimai buvo šalinami. Taigi į sąrašą buvo įtrauktos tik naujos nuorodos. Tai leido labai neapkrauti tinklo srauto.

Atlikus prototipo bandymus, kai yra pasirinkti du duomenų šaltiniai (interneto svetainės: http://www.skelbiu.lt ir



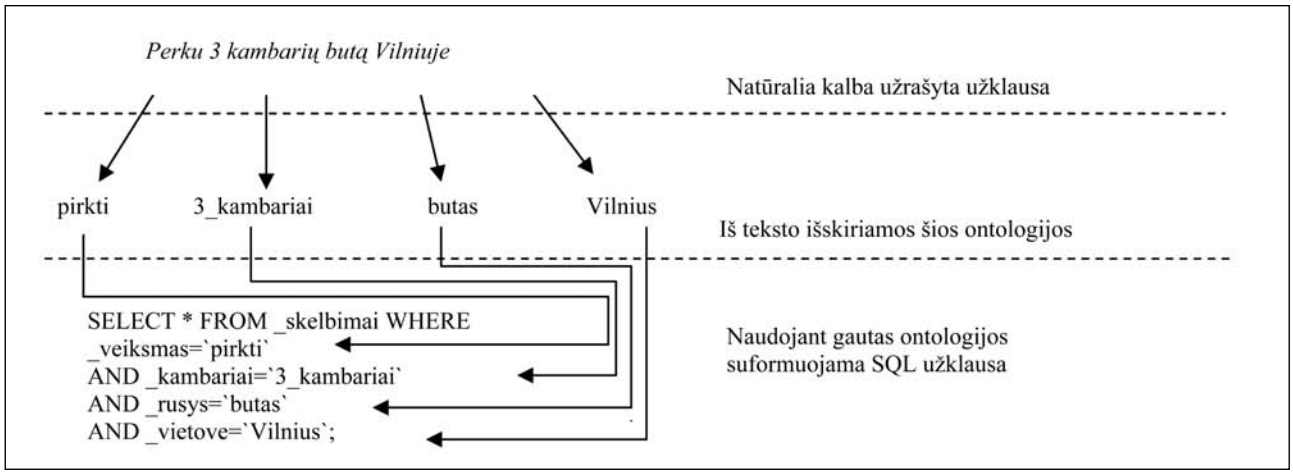
3 pav. Nekilnojamojo turto ontologijos kūrimas naudojant *Protege 3.2 beta* programinį įrankį

Fig 3. Application of Protégé 3.2. software for the realty ontology



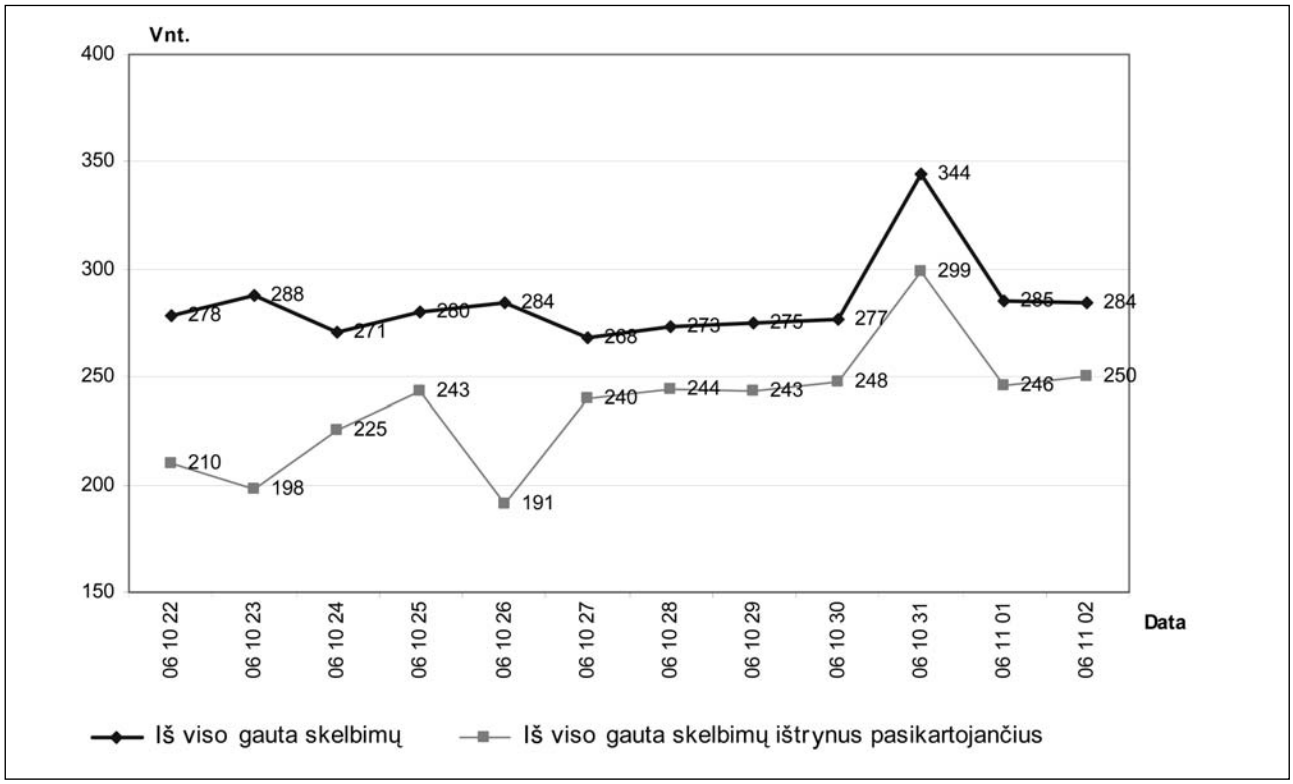
4 pav. Nekilnojamojo turto ontologijos naudojimas išgaunant kriterijus

Fig 4. Application of realty ontology for criteria retrieval



5 pav. Užklausos transformacija į SQL užklausą

Fig 5. Request transformation in SQL request



6 pav. Atliekamo eksperimento rezultatai

Fig 6. Results of the experiment

http://www.skelbimas.lt) yra gauti rezultatai (6 pav.). Visus atliekamus veiksmus ir procesų nuoseklumą kontroliuoja koordinuojantis agentas. Jo uždavinys – leisti tam tikru laiku duomenų išgavimo ir duomenų perkėlimo procesus. Duomenys išgaunami ir apdorojami du kartus per parą.

Pasikartojančių skelbimų vidurkis yra 17 %.

Buvo nustatyta, kad statistiką reikia skaičiuoti tą pačią dieną, nes skaičiuojami kitą dieną duomenys yra netikslūs (dėl šalinamų pasikartojančių skelbimų). Skaičiuoti statistikai buvo prataikyti MySQL 5.0 DBVS trigeriai. Statistika įrašoma į atskirą lentelę.

5. Išvados

Pasiūlytas sprendimas informacijai apie nekilnojamąjį turtą gauti iš įvairių interneto šaltinių, naudojant sukurtą nekilnojamojo turto ontologiją. Ši schema sėkmingai apsaugo gautus duomenis nuo pasikartojimo.

Nors aprašyta schema yra taikoma nekilnojamojo turto skelbimų paieškai, ją galima taikyti ir kitai informacijai ieškoti. Tai priklauso nuo naudojamos ontologijos.

Šio agento surinkti ir apdoroti duomenys gali būti sprendimo rengimo sistemos duomenų šaltinis, kuris gali atrinkti geriausią pasirinkimą iš siūlomo nekilnojamojo turto sąrašo.

Literatūra

1. DEAN, M.; SCHREIBER, G. *OWL Web ntology language reference.* W3C Recommendation. 10 February 2004. Available from Internet: <http://www.w3.org/TR/2004/REC-owl-ref-20040210>.
2. GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 1993, Vol 5, No 2.
3. SMITH, M. K.; WELTY, Ch.; McGUINNESS, D. L. *OWL Web ontology language guide*. W3C Recommendation. 10 February 2004. Available from Internet: <http://www.w3.org/TR/2004/ REC-owl-guide-20040210>.
4. CUNNINGHAM, H. *Information extraction – a user guide (Second Edition)*. Institute for Language, Speech and Hearing (ILASH), and Department of Computer Science University of Sheffield, UK, 1999. Available from Internet: <http://www.dcs.shef.ac.uk/~hamish>.
5. CUNNINGHAM, H.; MAYNARD, D. GATE: A framework and graphical development environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02*), 2002.
6. APPELT, D.; HOBBS, J. Fastus: Aa finite state processor for information extraction from real world text. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambery, France, August 1993.
7. DECKER, S.; ERDMANN, M. Ontobroker: ontology based access to distributed and semi-structured information. In R. Meersman et al., editors. *Database Semantics: Semantic Issues in Multimedia Systems*. Kluwer Academic Publisher, 1999, p. 351–369.
8. KIRYAKOV, A.; POPOV, B. *Semantic annotation, indexing, and retrieval*. Ontotext Lab, Sirma AI EOOD, Sofia, Bulgaria, 2005. 16 p.
9. KIRYAKOV, A.; POPOV, B. *Towards semantic web information extraction.* Ontotext Lab, Sirma AI EOOD, Sofia, Bulgaria, 2005. 21 p.
10. SCIME, A.; KERSCHBERG, L. *WebSifter: an ontological Web-mining agent for EBusiness*. 2003. 15 p.
11. SNOUSSIL, H. *Toward an Ontology-based Web Data Extraction*. 2005, Montréal, H3C 3J7 Canada. 8 p.

Darius JURKEVIČIUS. Dept of Information Systems of VGTU. Master's degree studies. Transport engineer (2005). Research interests: information systems engineering, information retrieval.

Algirdas LAUKAITIS. Graduate of Vilnius University, Faculty of Physics in 1992. PhD degree from the Institute of Mathematics and Informatics, Vilnius, 2002. Associated Professor of the Information Systems Department of Vilnius Gediminas Technical University. Research interests include text mining, natural language interfaces, machine translation systems and knowledge management.