Podpora adaptívneho WEB-u prostriedkami strojového učenia

Kristína Machová[1] Ivan Klimko[2]

[1]Katedra kybernetiky a umelej inteligencie, FEI, TU – Technická Univerzita Košice, Letná 9, 04200, Košice
Kristina.Machova@tuke.sk

[2]Katedra kybernetiky a umelej inteligencie, FEI, TU – Technická Univerzita Košice, Letná 9, 04200, Košice
I.Klimko@e-unicom.sk

Abstrakt. Príspevok sa zameriava na riešenie problému znižovania kognitívnej záťaže používateľa Internetu pomocou metód strojového učenia. Článok prezentuje systém AWS, ktorý bol navrhnutý tak, aby odporúčal internetové stránky používateľovi na základe modelu daného používateľa. Systém taktiež poskytuje informácie o modeloch návštevníkov pre účely správy obsahu servera. Systém AWS je sprievodcovský systém s off-line učením, s individuálnou adaptáciou a s podporou globálnej adaptácie obsahu servera.

Kľúčové slová: adaptívny web, strojové učenie, heuristicke prehľadávanie, zhlukovanie, model používateľa

Úvod

Internet je snáď najpoužívanejším informačným médiom dnešnej doby. Je preň charakteristický dynamický rozvoj a rýchly rast, čo so sebou prináša niektoré nevýhody. Jednou z nich je fakt, že Internet obsahuje obrovský počet stránok. Môže sa stať, že sa používateľ vzdá ďalšieho prehľadávania, aj keď hľadaná informácia je na stránkach prítomná. Ťažkosti spôsobuje aj hypertextovosť Internetu, spôsobená naakumulovaním množstva odkazov. Tieto problémy spojené s cieleným vyhľadávaním informácií sa merajú kognitívnou záťažou používateľa.
Cieľom predkladaného článku je prezentovať systém AWS, ktorý prispieva k znižovaniu kognitívnej záťaže používateľa Internetu. Je to systém zameraný na podporu adaptívneho web-u. Adaptívny web sa väčšinou automaticky prispôsobuje svojim návštevníkom na základe pozorovania ich správania sa. Tento spôsob práce vychádza z myšlienky inteligentných personálnych asistentov predstavenej v [4]. Existujú systémy pre adaptívny web, ktoré pracujú s technikami založenými na ohodnoteniach stránok používateľmi tzv. „collaborative filtering“ [6]. Používateľ dostáva odporúčania podľa toho o čo sa zaujímajú ostatní používatelia s rovnakým hodnotením, resp. s rovnakými či podobnými záujmami (portál Amazon.com). Iné systémy vykonávajú predikciu stránok (WebWatcher, AVANTI). Špecifickou kategóriou sú systémy inšpirované neurónovými sieťami, založené na Hebbovom učení [1].

Systém AWS sa zameriava na získanie modelu používateľa na základe požiadaviek používateľa. Tento model sa použije pri modifikácii odpovede používateľovi, ktorému sa doporučia iba dokumenty odpovedajúce jeho profilu. Model používateľa sa získava heuristickymi metódami strojového učenia.

Strojové učenie

Učenie modelu používateľa sa realizuje strojovým učením na základe protokolu www servera (IP návštevníka, jeho heslo, požiadavky návštevníka, cookies). Používané metódy strojového učenia sú uvedené v [3]. Jednotlivé prístupy na stránky servera (trénovacie príklady) sú charakterizované atribútom **A(url)**. Napríklad: **A(url)**=/som.php. Jednotlivé adresy stránok (URL) sú nahradené **kľúčovými slovami** popisujúcimi obsah stránok. (Popisy stránok boli vopred manuálne generované.) Napríklad /som.php=UI,NS,zhlukovanie,SOM,konkur_učenie. Tak vzniknú transformované trénovacie príklady (Tabuľka 1), používané systémom AWS na učenie modelov používateľov.

Tabuľka 1. Trénovacie príklady po transformácii

Por.č.	Kľúčové slová	T(userID)
1	UI,NS,zhlukovanie, SOM, konkur_učenie	USER3eafc6cd8c98a
2	UI, NS, konkur_učenie, MAXNET	USER3eafc6cd8c98a
3	UI, NS, dopredné, topológia	USER3eafc6cd8c98a
4	UI, SU, zhlukovanie, COBWEB	USER3eafc71274ad9
5	UI, SU, rozhodovanie, roz_stromy, ID3	USER3eafc71274ad9
6	UI, SU, rozhodovanie, roz_stromy, C4.5	USER3eafc71274ad9
7	Riadenie, regulátory, spojité, PID	USER3eafc7413857e
8	Riadenie, regulátory, diskrétne, PSD	USER3eafc7413857e
9	Riadenie, automaty, PLC	USER3eafc7413857e

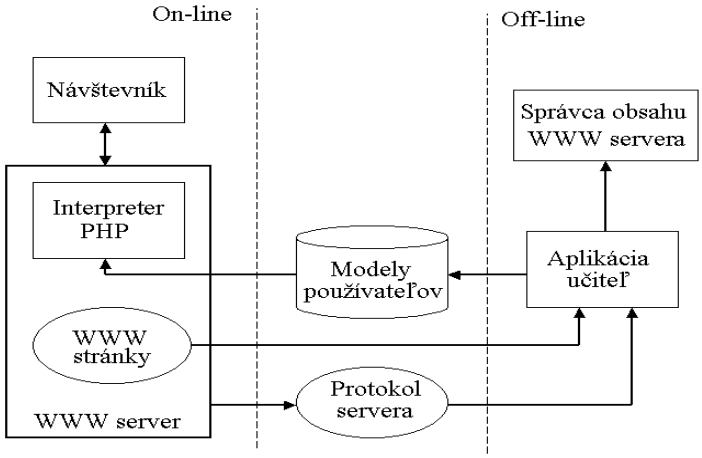
T(userID) identifikuje používateľa a tým udáva triedu, ku ktorej bude prístup na stránku zaradený. (Bolo by možné diskutovať o dostatočnosti učenia z prístupov používateľov na stránku, keďže návšteva stránky používateľom nemusí znamenať, že mu jej obsah vyhovuje. Avšak prístup na stránku, cez ktorú sa používateľ iba prekliká, tiež môže predstavovať užitočnú informáciu o nasmerovaní záujmov príslušného používateľa. Viac informácií o užitočnosti danej stránky pre používateľa by bolo možné získať meraním času, ktorý používateľ strávi štúdiom stránky. Prípadne vyžiadať od používateľa vyjadrenie, či je daná stránka pre neho zaujímavá, alebo nie.)

Klasifikácia je zaradenie nového príkladu (nepoznáme jeho triedu) do triedy **T** na základe definície triedy **T**, ktorá bola získaná niektorou metódou strojového učenia. V rámci systému AWS boli použité metódy heuristickeho prehľadávania priestoru kandidátov pojmov HGS (Heuristic General to Specific) a HSG (Heuristic Specific to General) [3], [4]. Tieto algoritmy redukujú priestor pojmov pomocou heuristickej ohodnocovacej funkcie aj pomocou Beam Size (**BS**). Líšia sa smerom prehľadávania priestoru pojmov.
Algoritmy HGS a HSG sú pre použitie v systéme AWS **modifikované** nasledovne: Algoritmy pracujú s čiastočným pokrytím a s premenným počtom atribútov. Za najvšeobecnejší popis sa berie úplná množina kľúčových slov, pretože uvažované kľúčové slová sú chápané v disjunktívnom zmysle. Operátorom špecifikácie (zovšeobecnenia) sa stáva operátor vypustenia (pridania) kľúčového slova z (do) popisu. Pri použití čiastočného pokrytia je potrebné definovať stupeň pokrytia. Stupeň pokrytia $\mathbf{P}_{\mathbf{pok}}$ je daný pomerom počtu kľúčových slov v príklade obsiahnutých v popise pojmu k počtu všetkých kľúčových slov v príklade. Samotná klasifikácia sa vykonáva za pomoci stupňa pokrytia. Stránka **S** je odporučená používateľovi **P** vtedy, ak stupeň pokrytia kľúčových slov stránky **S** a kľúčových slov modelu používateľa **P** je väčší ako zvolená hodnota prahu **H**. Taktiež je modifikovaná heuristická ohodnocovacia funkcia.
Problémom je, že v trénovacej množine sa prirodzene vyskytujú v značnom množstve protirečivé príklady. V aplikácii AWS sa používajú dva rozličné prístupy k tomuto problému: Prvým prístupom je implicitné riešenie modifikovanou heuristickou funkciou, ktorý je v ďalšom označovaný **„upravené heuristikou“**. V tomto prípade sa protirečivé príklady eliminujú. Druhý možný prístup - vymazanie protirečivých príkladov z množiny negatívnych príkladov, bude označovaný **„s vypustením“**.
Metódy **zhlukovania** (clustering) sa aplikujú v prípadoch, keď jednotlivé trénovacie príklady neobsahujú vopred informáciu o triede a teda trénovacie príklady sú rozdeľované do prirodzených skupín, resp. zhlukov. Hovoríme, že ide o techniky nekontrolovaného učenia, keďže neexistuje spätná väzba vo forme zadanej triedy. V rámci tejto práce bola použitá konceptuálna technika CLUSTER/2 [3],[5].

Popis systému AWS

Systém AWS bol navrhnutý tak, aby umožňoval odporúčanie stránok používateľovi na základe jeho modelu a tak uskutočňoval individuálnu adaptáciu. Model návštevníka je získaný učením z histórie, algoritmami HGS a HSG. Tieto algoritmy boli vybraté kvôli dobrej skúsenosti s nimi v rámci iných experimentov. Systém zároveň umožňuje aj poskytovanie informácií o modeloch návštevníkov a ich záujmoch pre účely správy obsahu servera. Zhlukovanie používateľov je v systéme realizované pomocou jednoduchej zhlukovacej techniky CLUSTER/2. Systém AWS podľa Obr. 1. pozostáva z on-line a off-line časti.
On-line časť je zodpovedná za identifikáciu návštevníka a následné generovanie odporúčaní podľa jemu zodpovedajúceho modelu. On-line časť je tvorená WWW serverom, interpretrom PHP a databázou modelov používateľov. Návštevník stránky zasiela požiadavky na WWW server. Server ich postúpi interpretru PHP. Interpreter

identifikuje používateľa a zadá dotaz do databázy modelov používateľov. Z nej vyzdvihne adresy a názvy príslušných odporúčaných stránok (podľa modelu daného používateľa) a vráti tieto odporúčania návštevníkovi spolu so žiadanou stránkou. (Odporúčania sa v momentálnom štádiu realizácie týkajú iba obsahu jedného servera.) Ako www server je použitý server Apache verzia 1.3.22, interpretrom PHP skriptov je PHP verzia 4.1.1. Použitá bola aj databáza MySql verzia 3.23.47. On-line časť je tvorená skriptami PHP tak, že ich do dokumentov stačí vložiť PHP direktívami include a require. Dokumenty však musia mať príponu php.



Obr. 1. Štruktúra systému AWS

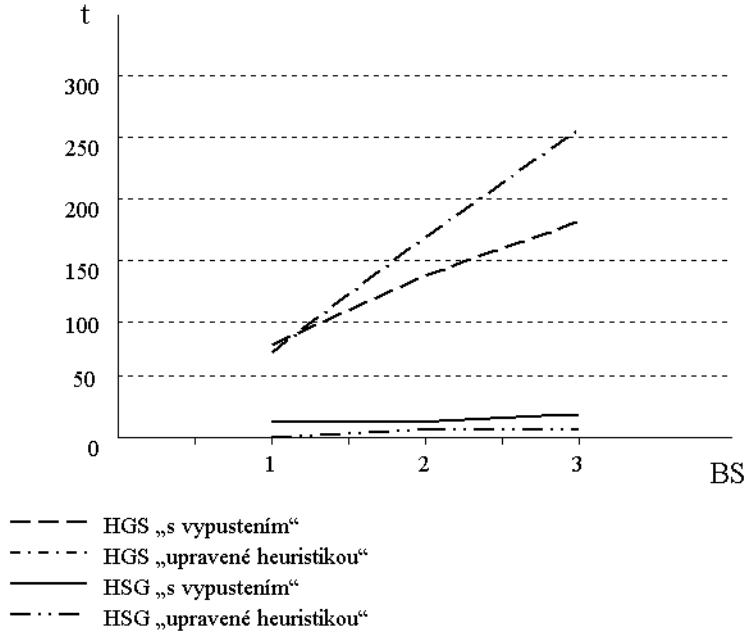
Off-line časť zodpovedá za vytváranie modelov používateľov a poskytovanie informácií pre adaptáciu obsahu servera. Učenie je realizované na základe informácií o obsahu servera a protokolu servera. Aplikácia pracuje s protokolom servera vo formáte NCSA Combined Log Format. Spoločnou časťou systému je databáza, ktorá uchováva modely používateľov a protokol servera s informáciami o spracovaných požiadavkách návštevníkov. Systém umožňuje identifikáciu návštevníka pomocou IP adresy alebo cookies. Identifikácia pomocou cookies sa javí ako výhodnejšia. Aj keď na druhej strane má použitie cookies na identifikáciu používateľa svoje nevýhody, napríklad bezpečnostné riziká. Nevýhodou tohto prístupu je taktiež skutočnosť, že pomocou cookies je možné identifikovať návštevníka iba keď sa hlási zo svojho stroja. Respektíve iný užívateľ sa môže stotožniť s používateľom, ktorého stroj práve používa. Off-line časť je nosnou časťou systému. Je reprezentovaná aplikáciou ASW/Učiteľ. Zabezpečuje vlastné učenie modelu používateľov, priradenie stránok k jednotlivým modelom používateľov a podporu pre správu obsahu servera zhlukovaním modelov používateľov. Takto získava definície zhlukov vo forme reťazca kľúčových slov. Zároveň vypočíta početnosť zastúpenia návštevníkov v jednotlivých zhlukoch, čím definuje primárne skupiny návštevníkov. Modifikáciu obsahu servera vykonáva správca na základe vlastného uváženia, po zhodnotení

klastrov generovaných systémom AWS. Môže vylúčiť stránky, o ktoré dlhšiu dobu nebol záujem. Taktiež môže rozšíriť o nové stránky tie oblasti, s ktorými rezonujú najvačšie zhluky používateľov. Obsah samotných dokumentov ani výzor informačného zdroja sa nemení. Off-line časť (aplikácia AWS/Učiteľ) je naprogramovaná v programovacom jazyku C++ prostredia Borland C++ Builder 6.

Testovanie

Systém AWS bol predbežne testovaný na množine 31 internetových stránok z oblasti umelej inteligencie a riadenia. Pri testoch sa pracovalo s 8 návštevníkmi, ktorí pristupovali na vybrané internetové stránky. Testovanie bolo rozdelené na dve fázy. Prvá fáza sa týkala testov súvisiacich s generovaním odporúčaní používateľovi, kde bola skúmaná časová náročnosť jednotlivých algoritmov a rozdiel medzi generovanými popismi. Druhá fáza sa týkala testov súvisiacich so zhlukovaním, kde sa skúmala vhodnosť generovaných popisov zhlukov a ich početnosť.

Časová náročnosť algoritmov HGS a HSG



Obr. 2. Porovnanie časovej náročnosti HGS a HSG

Časové testy boli realizované na počítači s procesorom AMD K6-2/400 MHz, 152 MB RAM s operačným systémom Windows 2000 Professional.
Pri porovnaní časovej náročnosti algoritmov HGS a HSG na Obr. 2. boli zohľadnené priemerné časy výpočtu v sekundách. Z grafu vyplýva, že HGS potrebuje oveľa viac času na nájdenie riešenia, keďže hľadané riešenia sú na nízkej úrovni všeobecnosti, na ktorú algoritmus HGS zostupuje dlhšie ako na ňu vystupuje algoritmus HSG. Z grafu taktiež vyplýva, že algoritmus HGS „upravené heuristikou“ je najpomalší a s narastajúcim BS nároky na čas prudko stúpajú. Pri HGS „s vypustením“ nároky na čas s narastajúcim BS stúpajú miernejším tempom. Rozdiely v čase spracovania HSG „upravené heuristikou“ a HSG „s vypustením“ sú nevýrazné a pre obidve verzie platí, že čas spracovania narastá zanedbateľne. Pri algoritme HGS pri vymazaní protirečivých príkladov z množiny negatívnych príkladov je redukovaný priestor prehľadávania, čo sa výrazne prejaví skrátením času potrebného na dosiahnutie výsledku. Pri HSG naopak vymazanie protirečivých príkladov z množiny negatívnych príkladov umožní algoritmu postupovať na vyššie úrovne všeobecnosti, čo sa prejaví miernym predĺžením času výpočtu.

Špecifikácie generované pomocou HGS a HSG

Nájdené riešenia pri použití HGS „upravené heuristikou“ ilustruje Tabuľka 2. a nájdené riešenia pri použití HSG „s vypustením“ ilustruje Tabuľka 3.

Tabuľka 2. Modely generované HGS „upravené heuristikou“.

Používateľ	Model
USER3eafc6cd8c98a	Konkur_učenie, MAXNET, NS, SOM, topológia
USER3eafc71274ad9	CN2, COBWEB, HCT, IWP, prahové_pojmy
USER3eafc7413857e	PLC, programovanie
USER3eafc77ec0dbc	CLUSTER/2, SOM, zhlukovanie
USER3eafc7999925 1	NEX, rozhod_zoznamy
USER3eafc7bca8371	Fuzzy_regulátor
USER3eafc7dd6b31 2	Moduly, programovanie, simatic, SLC
USER3eafc807499a2	C4.5, CN2, fuzzy_regulátor, NEX, PLC, roz_zoznamy, simatic

Tabuľka 3. Modely generované HSG „s vypustením“.

Používateľ	Model
USER3eafc6cd8c98a	**UI, zhlukovanie**, konkur_učenie, MAXNET, NS, SOM, topológia
USER3eafc71274ad9	(CN2), **UI, SU, rozhodovanie, roz_stromy, ID3,** COBWEB, HCT, IWP, prahové_pojmy
USER3eafc7413857e	**riadenie, regulátory, spojité, PID,** PLC, (programovanie)
USER3eafc77ec0dbc	**UI, SU**, CLUSTER/2, SOM, zhlukovanie
USER3eafc79999251	**UI, SU, rozhodovanie, roz_stromy, ID3,** NEX, rozhod_zoznamy
USER3eafc7bca8371	**riadenie, fuzzy, regulátory**, fuzzy_regulátor
USER3eafc7dd6b312	**riadenie, automaty**, moduly, programovanie, (simatic), SLC
USER3eafc807499a2	**UI, SU, rozhodovanie, roz_stromy, ID3,** C4.5, (CN2, fuzzy_regulátor), NEX, (PLC, roz_zoznamy), simatic

Tabuľka 3. obsahuje zvýraznené kľúčové slová (resp. kľúčové slová v zátvorke), ktorých výskyt je nadbytočný (resp. chýbajú) v porovnaní s najšpecifickejšími výsledkami, ktoré ilustruje Tabuľka 2. Modely používateľov generované HGS „upravené heuristikou“ sú presnejšie, pretože obsahujú menej kľúčových slov, teda záujmy používateľov sú lepšie vyhranené.
Z daných výsledkov vyplýva, že rádovo pomalší algoritmus HGS poskytuje presnejšie výsledky. Pre HGS sa najviac osvedčili hodnoty klasifikačného prahu: 0.33 a 0.50. Pre HSG sa najviac osvedčili hodnoty prahov 0.50 a 0.60.

Zhlukovanie pomocou CLUSTER/2

Vzhľadom na štruktúru používateľov boli zvolené počty zhlukov 2 a 4. Pre počet zhlukov 2 boli vygenerované nasledovné zhluky:
Zhluk 1 s popisom: automaty, fuzzy, fuzzy_regulátor, moduly, PID, PLC, programovanie, regulátory, riadenie, SLC, spojité. Do tohto zhluku boli zaradení traja používatelia s primárnym záujmom o riadenie.
Zhluk 2 s popisom: C4.5, CLUSTER/2, COBWEB, HCT, ID3, IWP, konkur_ucenie, MAXNET, NEX, NS, prahové_pojmy, roz_stromy, rozhod_zoznamy, rozhodovanie, simatic, SOM, SU, topológia, UI, zhlukovanie. Do tohto zhluku boli zaradení piati používatelia so záujmom o strojové učenie. Podobne pre 4 zhluky dostávame výsledky, ktoré ilustruje Tabuľka 4.

Tabuľka 4. Štruktúra zhlukov generovaných CLUSTER/2.

Zhluk	Popis zhluku	Počet návštevníkov
1	C4.5, CLUSTER/2, COBWEB, HCT, ID3, IWP, NEX, prahové_pojmy, rozhod_stromy, rozhod_zoznamy, rozhodovanie, simatic, SOM, SU, UI	4
2	Automaty, moduly, programovanie, riadenie, SLC	1
3	Fuzzy, fuzzy_regulátor, PID, PLC, regulátory, riadenie, spojité	2
4	konkur_učenie, MAXNET, NS, SOM, topológia, UI, zhlukovanie	1

Tieto výsledky sú menej nápomocné správcovi servera ako pri celkovom počte dvoch zhlukov. Prvý a štvrtý zhluk by bolo totiž možné zlúčiť do jedného zhluku, podobne druhý a tretí zhluk. Tieto závery podporuje aj slabá podpora počtu používateľov v druhom a štvrtom zhluku, čo ich predurčuje k zlúčeniu s väčšími zhlukmi.

Záver

V práci je uvedený návrh systému AWS, ktorý bol predbežne testovaný na pomerne malej množine internetových stránok. V budúcnosti by bolo vhodné overiť použiteľnosť tohto systému pre rozsiahlejšie informačné priestory. Podľa predbežného testovania algoritmov sa HGS „upravené heuristikou“ ukázal ako presnejší a HSG ako rýchlejší algoritmus. V budúcnosti, na učenie modelu používateľa z histórie, by bolo vhodné použiť novšie nástroje strojového učenia. Taktiež by bolo vhodné pri ďalších experimentoch použiť aj inú zhlukovaciu techniku, ktorá nevyžaduje zadanie pevného počtu zhlukov.
Zlepšenie výsledkov by sa dalo dosiahnuť aj použitím ontológie pre pohyb v usporiadanom priestore popisov pojmov. Rovnako by sa mohol zakomponovať systém pre automatické získavanie kľúčových slov z dokumentov. Fakt, že sa berú do úvahy aj stránky, cez ktoré sa používateľ iba prekliká, by mohol spôsobovať problémy pri zložitejšej štruktúre webu. Možnosť aplikácie systému nad zložitejšie štruktúrovaným webom, sa javí ako zaujímavý problém pre budúce bádanie.
Tento príspevok vznikol v rámci VEGA grantu MŠ SR č. 1/8131/01 „Znalostné technológie pre získavanie a sprístupňovanie informácií“.

Referencie

1. Benko, P.: Internet - heterogénny distribuovaný informačný systém <alf.fei.tuke.sk/publ/2202/InetHDIS.ps>. FEI TU Košice, 2002, 60s.
2. Langley, P.: *Elements of Machine Learning*. Morgan Kaufmann Publishers, Inc., 1996, San Francisco, California. ISBN 1-55860-301-8.

3. Machová, K.: *Strojové učenie. Princípy a algoritmy*. ELFA s.r.o., 2002, Košice. ISBN 80-89066-51-8.
4. Michalski, R.S.: Pattern Recognition as Rule-guided Inductive Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*, 1980, 349-361.
5. Michalski, R.S., Stepp, R.: Automated Construction of Classification: Conceptual Clustering Versus Numerical Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, No.5,* 1983, 219-243.
6. Pareschi, R., Borghoff, U.M.: *Information Technology for Knowledge Management*. Springer-Verlang, Berlin, Heidelberg, 1998, ISBN 3-540-63764-8.

Annotation:

The paper focuses on solving the problem of Internet users' cognitive load decrease based on machine learning methods. It presents the AWS system designed to suggest Internet pages to a user on the base of his/her model. The system also offers information about visitor models for the purpose of the server content management. The information can be used for the customization of server's content to users. The AWS system is an advisory system with off-line learning capabilities, individual adaptation, and the support of global server content adaptation.