Strojno učenje za pomoč pri preiskovanju WWW

Dunja Mladenić
Institut Jožef Stefan
Odsek za inteligentne sisteme
Jamova 39, 1000 Ljubljana, Slovenija
E-mail: Dunja.Mladenic@ijs.si
http://www-ai.ijs.si/DunjaMladenic

Machine learning for better Web browsing

The paper describes design of **Personal WebWatcher**, *the system that uses some of the known machine larning techniques to help user browsing World Wide Web. The system observes users of World Wide Web and suggests pages they might be interested in. It learns user interests from the pages requested by the user. The learned model of user interests is then used to suggest hyperlinks on new HTML-documents requested by and presented to the user via Web browser that enables connection to "proxy" eg. Netscape.*

1. Uvod

Z naraščanjem razpoložljivih virov elektronskih informacij, še posebej nehomogenih, porazdeljenih virov kot je World Wide Web (WWW), narašča tudi interes za orodja, ki bi pomagala pri hitrem in učinkovitem iskanju željenih informacij. Na to problematiko se navezuje vse večje število raziskav, ki izhajajo iz obstoječih področij zbiranja in filtriranja informacij, strojnega učenja in inteligentnih agentov. Tako se zadnja leta intenzivno razvijajo nove metode strojnega učenja za uporabo na tekstovnih bazah podatkov, ki kombinirajo metode strojnega učenja z metodami filtriranja informacij. Te nove metode omogočajo tudi analizo pristopov posameznih uporabnikov do tekstovnih podatkov ter prilagajanje odziva sistemov individualnim potrebam uporabnikov. Na primer, Armstrong in sod. [2] so razvili sistem WebWatcher, ki pomaga uporabnikom pri iskanju informacij na WWW. Osnovna ideja je, da uporabnik poda ključne besede za iskanje, sistem pa svetuje hiperpovezave na HTML-dokumentih in shranjuje uporabnikovo oceno o relevantnosti prikazanih strani. Balabanović in sod. [3] so razvili sistem, ki se uči preiskovati Internet namesto uporabnika. Sistem preiskuje WWW in po določenem času izbere najboljše strani, ki jih pokaže uporabniku v oceno. Ocena strani se uporabi za izboljšanje hevristike iskanja in ocenjevanja kvalitete strani. Pazzani in sod. [17] zbirajo ocene uporabnikov o pregledanih WWW dokumentih in se na osnovi njih učijo profila uporabnikovih interesov. Pregledane dokumente razdelijo v skupine, glede na vsebino dokumenta, nato za vsako skupino zgradijo model. Tako generirana množica modelov predstavlja profil uporabnikovih interesov. Mitchell in sod. [15] so razvili sistem, ki se poveže na elektronski koledar uporabnika in gradi množico pravil, ki opisujejo navade uporabnika pri razporejanju sestankov. Sistem na osnovi tako zgrajenih pravil in nekaj dodatnih informacij o posameznih udeležencih sestankov pomaga uporabniku pri razporejanju novih sestankov. Lang [12] je razvil sistem za filtriranje elektronskih novic, ki je, podobno kot prej omenjeni sistemi, zasnovan na uporabi strojnega učenja na tekstovnih dokumentih. Krulwich in Burkey [11] sta razvila sistem, ki pregleduje elektronsko oglasno desko z vprašanji uporabnikov in jih napoti k osebam, ki bi jim znale pomagati pri iskanju odgovora. Sistem analizira vsebino prispelih vprašanj in jih razporeja v vnaprej definirana vsebinska področja. Maes [14] opisuje inteligenten uporabniški vmesnik, ki se uči tako od uporabnika kot tudi od istovrstnih vmesnikov drugih uporabnikov. Kot primer tovrstnega vmesnika so razvili sisteme za filtriranje elektronskih sporočil, filtriranje elektronskih novic, razporejanje sestankov na elektronskem koledarju uporabnika in svetovanje pri iskanju razvedrilnih informacij (npr.filmi, glasba). Nekateri od teh sistemov uporabljajo prilagojene metode filtriranja informacij na vsebini dokumentov (npr. filtriranje elektronskih novic), drugi pa temeljijo na korelaciji med interesi različnih uporabnikov (npr. iskanje razvedrilnih informacij). Drummond in sod. [6] so razvili sistem, ki pomaga uporabnikom pri preiskovanju elektronske knjižnice programov. Njihov sistem zahteva od uporabnika ključne besede in na njih uporabi svojo bazo pravil. Pri tem predpostavi vnaprej podan, fiksni format zapisov v elektronski knjižnici. Etizioni in Weld [7] sta razvila integrirani vmesnik do Interneta, ki s kombinacijo UNIX-ove lupine in WWW dosega do Interneta. Njun sistem sprejme od uporabnika zastavljene cilje in dinamično zgradi primerno zaporedje Internet ukazov za doseganje zahtevanega cilja. Burke in sod. [4] so razvili sistem za podporo dostopa do distribuiranih virov tekstovnih informacij in pomoč uporabniku pri iskanju odgovorov v tekstovnih bazah podatkov, kot so na primer baze s 'pogostimi vprašanji in ustreznimi odgovori' (FAQ). Pri tem je uporabniški vmesnik zasnovan na uporabi naravnega jezika pri postavljanju vprašanj.

World Wide Web postaja en najpopularnejših virov informacij, ki zelo hitro narašča in priteguje vse več uporabnikov. Uporabniki do njega dostopajo s pomočjo računalnikov, kar pomeni, da lahko skoraj neopazno uporabimo računalnik za sledenje, shranjevanje in analizo uporabnikovih dejavnosti. To nam lahko omogoči, da pomagamo uporabniku pri iskanju informacij na WWW.

Opisani sistem Personal WebWatcher se zgleduje po delu na uporabniku prilagojenem iskanju po WWW [2], [9] in s tem povezanem delu na uporabi strojnega učenja pri izdelavi uporabniku prilagojenih sistemov in gradnji modelov iz tekstovnih dokumentov [10], [12], [15]. Osnovna ideja sistemov, ki se prilagajajo uporabnikovim individualnim potrebam, je da poskušajo uporabiti vsako interakcijo z uporabnikom kot učni primer.

2. Uporaba strojnega učenja na tekstovnih dokumentih

Ob načrtovanju sistema za pomoč uporabniku pri pregledovanju WWW se porajajo raznovrstna raziskovalna vprašanja. Pogledali bomo samo nekatera med njimi, ki se nanašajo na uporabo strojnega učenja na tem problemu. Seveda obstaja tudi veliko drugih problemov, kot so na primer načrtovanje komunikacije med uporabnikom in sistemom (inteligentnim agentom), zagotovitev privatnosti zbranih podatkov, stopnja do katere naj sistem vključi uporabnika v svoje delo, ...

Uporabo strojnega učenja na tekstovnih dokumentih lahko opazujemo v luči treh pomembnih raziskovalnih vprašanj: (1) predstavitev tekstovnih dokumentov, (2) obravnavanje velikega števila atributov in (3) izbira učnega algoritma.

(1) Večina sistemov za učenje na tekstovnih dokumentih uporablja t.i. metodo 'vreče besed' za predstavitev tekstovnih dokumentov. Pri tej metodi dokumente predstavimo z vektorjem besed. Metoda predpostavi nepomembnost vrstnega reda in sosednosti besed v dokumentu. Nekateri pristopi v vektorjih besed uporabijo samo informacijo o tem ali se posamezne besede pojavijo v dokumentu (n.pr. [2], [5], [13], [17]), drugi pa upoštevajo število pojavitev posameznih besed (n.pr. [1], [3], [10]). Drugi raziskovalci uporabljajo še dodatno informacijo, kot je pozicija besede v dokumentu [5] ali opazujejo n-terice besed [20].

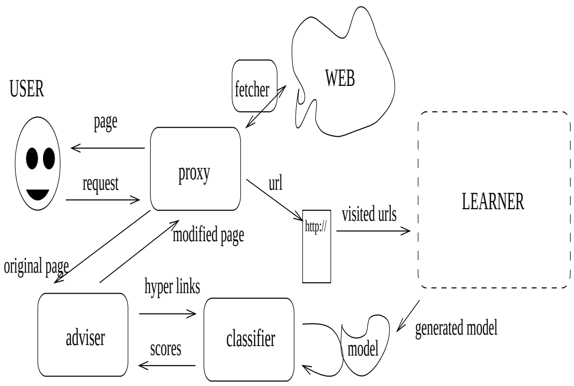
(2) Brisanje besed, ki se nahajajo na posebno v ta namen sestavljenem seznamu pogostih, neinformativnih angleških besed ali brisanje zelo pogostih oz. zelo redkih besed je eden od popularnih načinov za zmanjševanja števila atributov v tej domeni (n.pr. [5], [10]). V ta namen lahko uporabimo tudi metodo iz področja naravnega jezika, ki vsako besedo nadomesti z njenim korenom, je pa ta postopek seveda različen za različne jezike. Večina pristopov uporablja metodo obteževanja besed, ki je neodvisna od jezika dokumentov in omogoča sortiranje besed glede na uteži (ki odražajo njihovo pomembnost) in ohranjanje samo najboljših besed (n.pr. [1], [2], [3], [10], [13], [17]). Za zmanjševanje števila atributov (dimenzij prostora) nekateri uporabljajo tudi LSI metoda z dekompozicijo osnovnih vrednosti SVD ([8]).

(3) Pogosto uporabljana metoda učenja na tekstovnih dokumentih, ki izvira iz področja zajemanja in filtriranja informacij, je da vsak dokument predstavimo z vektorjem modificiranih frekvenc besed, ki se v njem pojavijo [19], naredimo vsoto vektorjev relevantnih dokumentov in tako dobljen sumarni vektor uporabimo pri klasifikaciji novih dokumentov (zasnovano na [18]). To metodo so nekateri že uporabili na WWW dokumentih in v večini primerov ugotovili, da daja slabše rezultate, kot metode strojnega učenja na istih prolemih (n.pr. [2], [3], [10], [17]).

Armstrong in sod. [2] so za gradnjo modela uporabili statističen pristop, ki predpostavlja medsebojno neodvisnost besed. Pazzani in sod. [17] so v svojih poskusih uporabili naivini Bayesov klasifikator, metodo najbližjega soseda in metodo odločitvenih dreves. Maes [14] tudi uporablja metodo najbližjih sosedov. Joachims [10] je predlagal modifikacijo omenjene metode iz področja zajemanja in filtriranja informacij, tako da upošteva tudi predstavitev dokumenta. Apté in sod. [1] so uporabili odločitvena drevesa. Cohen [5] uporablja odločitvena pravila in induktivno logično programiranje. Lewis in Gale [13] sta uporabila kombinacijo naivnega Bayesovega klasifikatorja in logistične regresije. McElligot in Sorensen [20] sta uporabila kombinacijo nevronskih mrež in genetskih algoritmov.

3. Personal WebWatcher

Personal WebWatcher je sestavljen iz dveh glavnih delov (glej sliko 1): **proxy serverja**, ki komunicira z uporabnikom preko bralnika za WWW dokumente in **dela za učenje**, ki gradi model uporabnikovih interesov. Prva verzija sistema vsebuje približno 2500 vrstic Perl kode (proxy) in 1500 vrstic C^{++} kode (strojno učenje).

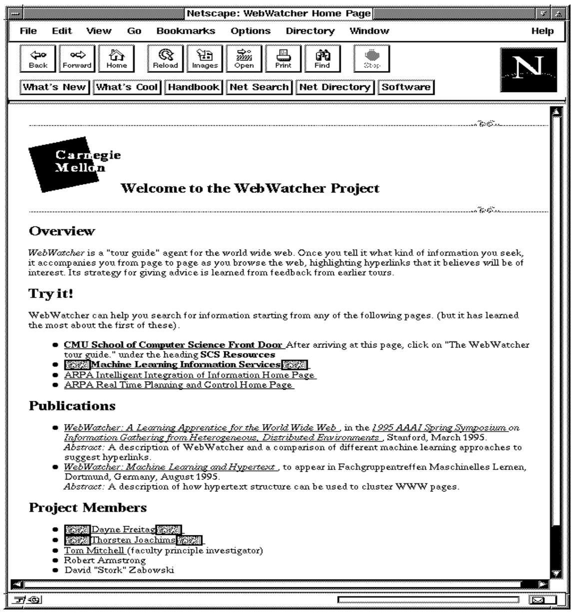


Slika 1: Struktura sistema Personal WebWatcher.

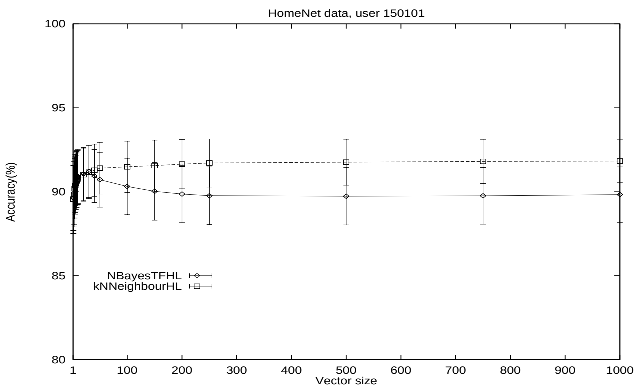
Proxy server je sestavljen iz treh glavnih delov: (1) **proxy** — uporablja zunanjo kodo za prenos zahtevanih strani **fetcher**, (2) **adviser** — pregleda trenutno stran in ji doda nasvete na osnovi rezultatov klasifikacije (3) **classifier** — klasificira nove hiperpovezave na osnovi modela uporabnikovih interesov (strojno učenje). Sistem deluje tako, da server čaka na zahtevo bralnika za WWW dokumente (eg. Netscape). Ob zahtevi prenese dokument, HTML-dokumentom doda nasvete in pošlje zahtevani dokument bralniku. Na primer, na sliki 2 je Personal WebWatcher svetoval (označil s kvadratkom) tri hiperpovezave: "Machine Learning Information Services" in dve domači strani raziskovalcev (Dayne Freitag, Thorsten Joachims). Na vrhu dokumenta je tudi dodana črta, kot informacija da je Personal WebWatcher aktiven in "opazuje uporabnika".

Model uporabnikovih interesov uporabimo za napovedovanje ali je novi dokument primer dokumenta, ki ustreza interesom uporabnika. V našem primeru uporabnik čaka na zahtevani dokument medtem, ko sistem analizira vsebino dokumenta in klasificira (napoveduje zanimivost) hiperpovezav na dokumentu. Model uporabnikovih interesov gradimo n.pr. vsako noč z uporabo strojnega učenju. Pri tem kot učne primere uporabimo hiperpovezave z dokumentov, ki smo jih pokazali uporabniku. Primere razdelimo v dve množici (pozitivne in negativne), glede na to, ali je uporabnik zahteval dokument, na katerega kaže hiperpovezava. V prvih poskusih smo za gradnjo modela ($User_H L$) uporabili naivni Bayesov klasifikator in metodo k-najbližjih sosedov $User_{HL} : HyperLink \to \{pos, neg\}$.

Sistem smo evaluirali na realnih podatkih uporabnikov WWW. Pri tem smo opazovali vpliv števila uporabljenih najinformativnejših atributov na rezultate prečnega testiranja (10-fold cross validation) in primerjali delovanje dveh metod strojnega učenja. Po rezultatih prvih poskusov kaže, da metoda



Slika 2: Primer HTML-strani prikazane uporabniku Personal WebWatcherja.



Slika 3: Vpliv števila atributov (vector size) na klasifikacijsko točnost za enega od uporabnikov.

k-najbližjih sosedov dosega nekoliko boljšo klasifikacijsko točnost kot naivni Bayesov klasifikator. Na sliki 3 so zaradi prostorskih omejitev podani rezultati pri enemu od uporabnikov. Bolj natančen opis sistema in več rezultatov je podano v [16]. V obeh primerih je klasifikacijska točnost približno enaka točnosti, ki bi jo dobili z napovedovanjem večinskega razreda (čez 90% dokumentov je ocenjenih kot nezanimivi za uporabnika) in je že po 100-200 atributih na svoji maksimalni vrednosti. V nadaljevanju dela se bomo najprej usmerili na opazovanje vpliva metode za izbiro atributov (v našem primeru besed) na rezultate sistema in možnosti evaluacije še z drugimi metodami (n.pr. informacijska vsebina odgovora), raziskovanje možnosti za bogatejšo predstavitev dokumentov, raziskovanje vpliva števila primerov na rezultate sistema in obravnavanje sprememb interesov uporabnika. Seveda je še mnogo drugih zanimivih raziskovalnih vprašanj iz področja strojnega učenja na velikih, dinamičnbih tekstovnih bazah podatkov in iz področja inteligentih agentov, ki bi jih v nadaljnem delu želeli raziskati.

Literatura

[1] Apté, C., Damerau, F., Weiss, S.M., Toward Language Independent Automated Learning of Text Categorization Models, *Proc. of the 7th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dubline, 1994.

[2] Armstrong, R., Freitag, D., Joachims, T., Mitchell, T., WebWatcher: A Learning Apprentice for the World Wide Web, *AAAI 1995 Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, Stanford, March 1995.

[3] Balabanović, M., Shoham, Y., Learning Information Retrieval Agents: Experiments with Automated Web Browsing, *AAAI 1995 Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, Stanford, March 1995.

[4] Burke, R., Hammond, K., Kozlovsky, J., Knowledge-based Information Retrieval for Semi-Structured Text, *Working Notes from AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, pp. 19–24, American Association for Artificial Intelligence, 1995.

[5] Cohen, W.W., Learning to Classify English Text with ILP Methods, *Workshop on Inductive Logic Programming*, Leuven, September 1995.

[6] Drummond,C., Ionescu, D., Holte, R., A Learning Agent that Assists the Browsing of Software Libraries, *Technical Report TR-95-12*, Computer Science Dept., University of Ottawa, 1995.

[7] Etizioni, O., Weld, D., A Softbot-Based Interface to the Internet, *Communications of the ACM* Vol. 37, No. 7, pp.72–79, July 1994.

[8] Foltz, P. W. and Dumais, S. T., Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, 35(12), pp.51—60, 1992.

[9] Joachims, T., Mitchell, T., Freitag, D., Armstrong, R., WebWatcher: Machine Learning and Hypertext, *Fachgruppentreffen Maschinelles Lernen*, Dortmund, August 1995.

[10] Joachims, T., A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, *Technical report, CMU-CS-96-118*, School of Computer Science, Carnegie Mellon University, March 1996.

[11] Krulwich, B., Burkey, C., The ContactFinder agent: Answering buletin board questions with referrals, *Proceedings of the Thirteenth National Conference on Artificial Intelligence AAAI 96*, pp.10—15, 1996.

[12] Lang, K., News Weeder: Learning to Filter Netnews, *Proc. of the 12th International Conference on Machine Learning ICML95*, 1995.

[13] Lewis, D.,D., Gale, W., A., A Sequential Algorithm for Training Text Classifiers, *Proc. of the 7th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dubline, 1994.

[14] Maes, P., Agents that Reduce Work and Information Overload, *Communications of the ACM* Vol. 37, No. 7, pp.30–40, July 1994.

[15] Mitchell, T., Caruana, R., Freitag, D., McDermott, J., Zabowski, D., Experience with a Learning Personal Assistant, *Communications of the ACM* Vol. 37, No. 7, pp.81–91, July 1994.

[16] Mladenić, D., Personal WebWatcher: Implementation and Design. *IJS DP-7472*, Ljubljana, Slovenija, 1996.

[17] Pazzani, M., Muramatsu, J., Billsus, D., Syskill & Webert: Identifying interesting web sites, *AAAI Spring Symposium on Machine Learning in Information Access*, Stanford, March 1996 and *Proceedings of the Thirteenth National Conference on Artificial Intelligence AAAI 96*, pp.54—61, 1996.

[18] Rocchio, J., Relevance Feedback in Information Retrieval, in *The SMART Retrieval System: Experiments in Automatic Document Processing*, Chapter 14, pp.313—323, Prentice-Hall Inc., 1971.

[19] Salton, G., Buckley, C., Term Weighting Approaches in Automatic Text Retrieval, *Technical report, COR-87-881*, Department of Computer Science, Cornell University, November 1987.

[20] Sorensen, H., McElligott, M., PSUN: A Profiling System for Usenet News, *CIKM'95 Intelligent Information Agents Workshop*, Baltimore, December 1995.