Portal web de ştiri autonom bazat pe prelucrarea limbajului natural

Traian Rebedea
Universitatea „Politehnica” Bucuresti, Facultatea de Automatică şi Calculatoare
Splaiul Independenţei, Nr. 313, Bucureşti, România
trebedea@gmail.com

Costin-Gabriel Chiru
Universitatea „Politehnica” Bucuresti, Facultatea de Automatică şi Calculatoare
Splaiul Independenţei, Nr. 313, Bucureşti, România
chirucos@gmail.com

Ştefan Trăuşan-Matu
Universitatea „Politehnica” Bucuresti, Facultatea de Automatică şi Calculatoare
Splaiul Independenţei, Nr. 313, Bucureşti, România
Institutul de Cercetare în Inteligența Artificială al Academiei Române,
Calea 13 Septembrie, Nr 13, Bucureşti, România
trausan@cs.pub.ro

REZUMAT
În această lucrare este prezentat un portal web de ştiri pentru limba română care încorporează un modul autonom de grupare şi de clasificare a ştirilor. Pentru a asigura funcționarea complet autonomă a portalului sunt folosite tehnici de prelucrare a limbajului natural. Ştirile sunt colectate automat de la un număr însemnat de surse folosind sindicalizarea web. Apoi, sunt utilizate tehnici de învățare automată pentru clasificarea ştirilor. În primul rând, acestea sunt grupate folosind un algoritm aglomerativ, iar cele mai numeroase grupuri corespund celor mai importante subiecte ale momentului. În acest fel se asigură şi colectarea mai multor informații despre aceste subiecte, întrucât sunt utilizate ştiri din diverse surse. În al doilea rând, portalul foloseşte un algoritm de clasificare automată pentru a încadra fiecare grup de ştiri într-un număr predeterminat de categorii. Atât pentru testarea, cât şi pentru evaluarea performanțelor clasificatoarelor au fost folosite peste o mie de ştiri pre-etichetate. Lucrarea conține şi o comparație completă a rezultatelor obținute pentru algoritmii de clasificare.

Cuvinte cheie
Portal de ştiri, Agent inteligent, Grupare, Clasificare, Prelucrarea limbajului natural

Clasificare ACM
H5.2. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCERE

În ultimii 10 ani am asistat la o adevărată explozie a web-ului, atât ca număr de domenii web existente, cât şi ca trafic şi număr de utilizatori. Avansul tehnologic din această perioadă a facilitat accesul la Internet al unui număr mare de utilizatori, oferindu-le acestora viteze de transfer din ce în ce mai mari. Studiile Internet World Stats [5] relevă faptul că în ultimii 5 ani numărul de utilizatori de Internet la nivel mondial s-a dublat, ajungând la 20% din populația lumii – adică peste 1 miliard de utilizatori. Creşterea numărului de utilizatori (consumatori) a determinat intrarea pe această piață a unui număr din ce în ce mai mare de ofertanți. Astfel, conform datelor oferite de Netcraft [7] – vezi figura 1, tot în ultimii 5 ani, numărul total al siturilor web s-a triplat, ajungându-se în februarie 2008 la aproximativ 158 milioane de situri web, dintre care aproape jumătate sunt active – evoluția siturilor active fiind similară în această perioadă cu cea a siturilor totale. În aceste condiții, nu este o surpriză că volumul de informație disponibil pe web a crescut exponențial în această perioadă, ajungând la cifre impresionante. Yahoo a făcut public în august 2005 [14] că indexează peste 20 de miliarde de articole, dintre care 19,2 miliarde de documente web, 1,6 miliarde de imagini şi peste 50 de milioane de fişiere audio şi video. Numărul mare de documente aflate pe web, precum şi starea de continuă schimbare a acestuia pot să convertească web-ul într-un sistem haotic. Astfel, ca o consecință directă a popularității de care se bucură ca sistem global de informare, web-ul este bombardat în mod constant de cantități mari de date şi informație, care fac din găsirea informației utile pe web o experiență dificilă şi frustrantă. Soluția pentru găsirea informației pe web sunt motoarele de căutare. Odată cu redefinirea importanței unei pagini web de către Brin şi Page [8], motoarele de căutare au intrat într-o nouă etapă, oferind rezultate satisfăcătoare pentru majoritatea căutărilor. Totuşi, odată cu dezvoltarea web-ului, rezultatele unei căutări, chiar dacă este bine formulată, oferă câteva sute de rezultate utilizatorului.

Volumul mare de informații nu este singura problemă indusă de dezvoltarea web-ului. O altă problemă importantă este redundanța informațională întrucât multe dintre aceste informații se repetă în surse diferite. Pentru a oferi rezultate mai bune utilizatorului se pot crea programe care să exploateze exact aceste caracteristici – prin procesarea cantității mari de informație existentă, precum şi prin determinarea şi prelucrarea datelor redundante se pot obține rezultate folositoare.

Ideea principală a acestei lucrări este de a folosi redundanța informațională din cadrul fluxurilor de ştiri preluate sub formă de sindicalizare web. Pentru aceasta se folosesc metode avansate de procesare a textului, incluzând gruparea şi clasificarea, cu scopul de a obține un portal de ştiri complet autonom. Lucrarea continuă cu o secțiune în care este introdusă ideea de prelucrare inteligentă a ştirilor şi a conceptelor sale fundamentale. A treia secțiune descrie cele mai importante tehnici folosite pentru gruparea şi clasificarea textelor. Secțiunea

următoare conţine descrierea portalului web de ştiri pentru limba română ce foloseşte prelucrări inteligente, precum şi rezultatele obţinute. Lucrarea se încheie cu concluzii şi referinţe.

PROCESAREA INTELIGENTĂ A ŞTIRILOR

Cele mai importante surse de ştiri din lume au introdus sindicalizarea web ca o modalitate nouă de furnizare a ştirilor către cititori sau către alte situri web sau aplicaţii care doresc să le folosească. Formatele bazate pe XML folosite pentru sindicalizarea web au fost proiectate astfel încât conţinutul lor să fie uşor de folosit de către programe numite cititoare de flux de ştiri sau agregatoare. Acestea automatizează procesul de colectare periodică a fluxurilor de ştiri şi de prezentare a acestora utilizatorului într-o manieră uşor de urmărit. Însă nu rezolvă cele două probleme esenţiale: cantitatea mare de informaţie şi redundanţa ştirilor. Într-o fază iniţială a prelucrării fluxurilor de ştiri, s-au dezvoltat portaluri web de ştiri care au folosit numărul de cititori al unei ştiri drept criteriu de calcul al importanţei ştirii. Deşi prezintă unele avantaje clare – printre care determinarea importanţei în funcţie de preferinţele utilizatorilor, această metodă are şi multe dezavantaje: nu rezolvă nici problema cantităţii mari de informaţie, nici pe cea a redundanţei ştirilor, şi nu oferă un criteriu automatizat pentru calculul importanţei a priori a unei ştiri – înainte de a fi citită de cineva. În plus, fiecare ştire este legată de sursa din care a fost preluată, fără a se oferi surse alternative în legătură cu acel subiect. Alternativa propusă în această lucrare doreşte rezolvarea tuturor acestor probleme.

În ultima perioadă, mai multe echipe de cercetare, precum şi companii de renume şi-au axat eforturile spre prelucrarea inteligentă a fluxurilor de ştiri. De ce numesc aceste prelucrări inteligente? In primul rând, rezultatul acestor prelucrări determină automat subiectele zilei şi oferă o clasificare a acestor subiecte în funcţie de numărul de ştiri ce le alcătuiesc, prezentând în acest fel utilizatorului ştirile cu adevărat importante. În acest caz, determinarea importanţei unui subiect are la bază nu opţiunea subiectivă a cititorilor, ci pe cea *"obiectivă"* a specialiştilor, a agenţiilor de ştiri şi a ziarelor ce relatează diversele subiecte. În al doilea rând, pentru realizarea acestor prelucrări se utilizează o serie de tehnici de învăţare automată.

Ştirile preluate din cadrul diverselor fluxuri pot fi folosite pentru două prelucrări esenţiale: prima este *gruparea („clustering")* ştirilor şi aflarea subiectelor celor mai importante, iar a doua o constituie clasificarea automată a ştirilor în diverse categorii de ştiri predefinite. În plus, unele lucrări din domeniul extragerii de informaţii din ştiri [11, 12] propun o prelucrare mai amănunţită a acestora, cu accentul pe extragerea de cunoştinţe specifice – precum nume de persoane, de ţări, de organizaţii şi de companii – şi folosirea acestor cunoştinţe pentru gruparea şi clasificarea ştirilor. Desigur că există o multitudine de abordări în domeniul prelucrării inteligente a ştirilor, majoritatea aflându-se încă în faza de cercetare. O idee interesantă o reprezintă asignarea de etichete (cuvânt sau grup de cuvinte cheie, foarte semnificative) fiecărui grup de ştiri rezultat din grupare. În acest fel, utilizatorul poate urmări mult mai uşor lista de subiecte relativ la o anumită etichetă. Problema esenţială a acestei abordări o reprezintă calitatea etichetării, întrucât nu s-au găsit algoritmi de etichetare care să ofere rezultate satisfăcătoare [11]. O metodă nouă este oferită de sistemul *NewsJunkie* [4] ce propune determinarea noutăţii informaţiei unei ştiri – astfel având un flux de ştiri despre un subiect comun relatat de-a lungul unei perioade mai lungi de timp, sistemul îşi propune găsirea informaţiilor cu adevărat noi şi filtrarea articolelor ce relatează evenimente deja prezentate. O idee similară este propusă şi în *Ranking a Stream of News* [2].

Gruparea ştirilor exploatează redundanţa informaţiei obţinute prin colectarea unui număr mare de ştiri din diverse surse, obţinându-se grupuri de ştiri cu subiect comun. Redundanţa informaţiilor dintr-o colecţie de ştiri poate fi utilizată spre atingerea a două scopuri. În primul rând, poate fi folosită drept un criteriu de calcul al importanţei unei ştiri/subiect pornind de la ideea că numărul de apariţii ale unui subiect în surse diferite este proporţional cu importanţa acestuia. În al doilea rând, ştirile pot fi grupate în jurul acestor subiecte, reducând cantitatea de informaţie prezentată utilizatorului. În plus, această modalitate oferă avantajul de a prezenta fiecare subiect relatat din perspectiva fiecărei surse, permiţându-i cititorului urmărirea subiectului din sursa dorită sau compararea relatărilor din mai multe surse. Importanţa unei ştiri este determinată de către comunitatea media a agenţiilor de presă – această măsură putând fi considerată profesionistă şi obiectivă.

Clasificarea automată a ştirilor poate fi folosită cu succes împreună cu gruparea ştirilor, descrisă anterior, pentru crearea unui portal de ştiri care să funcţioneze complet automatizat şi să ofere utilizatorului toate facilităţile unui portal de ştiri obişnuit. Toate siturile de ştiri oferă o structurare a ştirilor în categorii pentru a furniza utilizatorilor informaţiile necesare în funcţie de interes. Chiar dacă standardele folosite pentru sindicalizarea web oferă posibilitatea definirii categoriei unui articol, faptul că nu există o taxonomie de ştiri valabilă universal, face ca această facilitate să fie practic inutilă în cadrul unui portal care culege ştiri de la un număr mare de surse. Singura opţiune este de a se folosi algoritmi de clasificare pentru asignarea fiecărei ştiri la o anumită categorie. Astfel, într-o primă etapă, se vor aduna ştiri care vor fi asignate manual sau pseudo-automat fiecărei categorii definite de portal. Întrucât există sindicări web ce conţin ştiri dintr-un singur domenii, procesul de colectare a ştirilor necesare pentru antrenare poate fi automatizat. Rezultatele obţinute folosind diverşi algoritmi de clasificare sunt promiţătoare, permiţând folosirea acestei tehnici cu rezultate foarte bune [3]. Deoarece ştirile sunt grupate în funcţie de subiect, algoritmul de clasificare poate fi folosit pentru grupuri şi nu pentru fiecare ştire individual – fiecare ştire din cadrul unui grup aparţinând categoriei în care este încadrat grupul respectiv.

TEHNICI DE GRUPARE ŞI CLASIFICARE A TEXTELOR

Metodele de grupare şi clasificare sunt utilizate des în prelucrarea limbajului natural. O ştire poate fi reprezentată ca un vector, ale cărui elemente sunt asociate cuvintelor cheie din text. Valorile elementelor pot fi şi variabile booleene reprezentând apariţia sau lipsa unui cuvânt cheie

din textul asociat ştirii, şi anume titlul şi descrierea preluate din sindicalizarea web. O îmbunătăţire a acestei reprezentări este de a folosi frecvenţa fiecărui termen sau măsura TF-IDF [1, p. 56-57]. Atât algoritmii de grupare, cât şi cei de clasificare necesită calcularea similarităţii între două elemente, în acest caz, între documente text. Spaţiul caracteristicilor folosit pentru reprezentarea documentelor are dimensiunea egală cu numărul de termeni distincţi din toate documentele, indiferent de metodele prezentate anterior. Transformarea distanţelor metrice în măsuri de similaritate nu oferă rezultate bune pentru texte. De aceea, pentru calculul similarităţii între documentele text sunt folosite alte măsuri [10], iar una dintre cele mai utilizate este cosinusul unghiului format între vectorii ce corespund documentelor care sunt comparate.

Gruparea datelor [6, p. 495-528] este o tehnică de analiză statistică a datelor, folosită cu succes în diverse subdomenii ale inteligenţei artificiale, precum învăţarea automată, mineritul datelor, recunoaşterea formelor şi analiza imaginilor. Gruparea este o modalitate de partiţionare a datelor în subseturi – grupuri, *clustere* – astfel încât datele din fiecare grup să aibă o caracteristică comună. De obicei, gruparea datelor se face în funcţie de proximitatea elementelor de grupat, pentru aceasta folosindu-se o funcţie distanţă. Gruparea datelor poate fi privită ca o tehnică de generalizare, întrucât datele de intrare sunt împărţite în clase de echivalenţă folosind anumite criterii. Toate elementele ce formează o clasă de echivalenţă moştenesc atributele grupului din care fac parte.

Algoritmii de grupare pot fi împărţiţi folosind mai multe criterii [10]. O prima clasificare este în algoritmi divizivi („*top-down*”) şi algoritmi aglomerativi („*bottom-up*”). Procesarea top-down porneşte cu toate datele într-un singur grup, pe care apoi îl rafinează în subgrupuri. Algoritmii aglomerativi, consideră fiecare element ca un grup separat, iar acestea sunt grupate ulterior. În funcţie de rezultatul grupării, avem de a face cu algoritmi ierarhici şi algoritmi de tip centroid („*flat*”). Algoritmii ierarhici oferă ca rezultat o structură arborescentă ce poate fi vizualizată ca o dendrogramă. De asemenea, gruparea poate fi tare („*hard*”) sau slabă („*soft*”). Gruparea tare va produce clustere ce nu se întrepătrund, adică fiecare element aparţine unui singur grup. Gruparea slabă asignează fiecărui obiect o probabilitate de a fi membrul unui grup.

Algoritmii ierarhici folosesc aproape întotdeauna o grupare tare, numai cei de tip centroid abordând şi gruparea slabă. Algoritmii de grupare ierahici sunt algoritmi greedy, ce folosesc pentru construirea arborelui rezultat ori o abordare top-down, ori una bottom-up. Algoritmii divizivi pornesc cu toate elementele considerate într-un singur cluster şi la fiecare pas, se determină grupul cel mai puţin coerent, acesta urmând ales să fie divizat. Grupurile ce conţin elemente mai similare au o coerenţă mai mare decât grupurile ce conţin elemente mai puţin similare. Algoritmii aglomerativi încep cu fiecare element ca un cluster separat şi, la fiecare pas, se aleg cele mai similare două clustere spre a fi unite. În cadrul algoritmilor ierarhici aglomerativi există trei metode de determinare a similarităţii dintre două grupuri – pentru a afla ce grupuri trebuie unite la fiecare pas. Prima abordare, defineşte similaritatea între două grupuri ca similaritatea între cei mai similari doi membri, câte unul din fiecare grup. Această variantă se numeşte legătură simplă („*single link*”) şi corespunde unui algoritm de determinare a arborelui minim de acoperire pentru setul de puncte determinate de elementele ce trebuie grupate. A doua abordare, denumită legătură completă („*complete link*”), defineşte similaritatea a două grupuri ca similaritatea celor mai puţin similare două elemente ce aparţin unul fiecărui grup. Această abordare produce rezultate mult mai bune decât cea anterioară, întrucât foloseşte calitatea globală a unui grup şi nu pe cea locală – prima abordare are tendinţa de a produce clustere „alongate”. Dezavantajul metodei legătură completă îl reprezintă complexitatea mai mare decât cea a metodei legătură simplă. Pentru a păstra performanţele metodei legătură completă şi complexitatea redusă a metodei legătură simplă, se foloseşte o abordare ce defineşte similaritatea a două grupuri ca fiind similaritatea medie între elementele fiecărui grup în parte.

Clasificarea automată este definită ca procesul de asignare a unor obiecte din universul problemei în două sau mai multe categorii sau clase. Procesul de clasificare statistică poate fi împărţit în două etape distincte: etapa de antrenare a clasificatorului şi etapa de utilizare efectivă a acestuia. Pentru etapa de antrenare, datele de intrare sunt numărul de categorii distincte şi elemente ce aparţin fiecărei categorii (un element poate aparţine uneia sau mai multor categorii). Aceste obiecte formează setul de antrenament şi poate fi reprezentat ca un set de (x, c) unde x este un vector m-dimensional ce codifică un obiect din setul de antrenament, iar c este clasa căreia obiectul îi aparţine. Pentru fiecare algoritm se defineşte o procedură de antrenare al cărei rezultat este un model de date (clasificatorul) pe baza căruia se pot clasifica obiectele din universul problemei. Acest model este implicat în cadrul etapei de utilizare efectivă a clasificatorului. Algoritmii de clasificare diferă prin modul cum este definită procedura de antrenare, deci şi prin modelul folosit pentru clasificator.

Există două categorii importante în care se împart algoritmii de clasificare. Prima variantă foloseşte o separare a universului problemei, care este de obicei un spaţiu vectorial multidimensional, în regiuni diferite, fiecărei regiuni fiindu-i asignată câte o categorie. Aceşti algoritmi nu folosesc praguri de încredere sau clase de probabilitate şi în această categorie se află algoritmul NN („*Nearest Neighbour*”) şi algoritmul de selecţie a caracteristicilor. A doua categorie priveşte problema clasificării ca pe o problemă de estimare a unei funcţii de forma $P(clasa \mid \bar{x}) = f(\bar{x}, \bar{\theta})$, unde x este vectorul de intrare, iar funcţia f este de obicei parametrizată cu vectorul $\bar{\theta}$. În acest caz, algoritmul de clasificare trebuie să determine din familia de funcţii parametrizate funcţia ce clasifică cel mai bine setul de antrenament – deci avem o problemă de optimizare. Aceşti algoritmi folosesc un model matematic mai complex, dar oferă rezultate mult mai bune. Printre aceştia se numeră reţelele Bayes şi maşinile vectorilor de suport („*SVMs – Support Vector Machines*”).

Categorizarea documentelor text [1, p. 124-169] este una dintre aplicaţiile algoritmilor de clasificare în domeniul prelucrării limbajului natural, având scopul de a clasifica textele în funcţie de subiect. Algoritmii de clasificare sunt folosiţi de o perioadă lungă de timp în inteligenţa automată, recunoaşterea formelor şi mineritul (bazelor) de date; trebuie avut însă în vedere că în toate aceste domenii datele sunt mult mai „structurate" decât în cazul textului. Una dintre caracteristicile definitorii pentru texte este numărul extrem de mare de termeni, de ordinul miilor sau zecilor de mii, spre deosebire de mineritul datelor unde numărul de coloane poate fi de ordinul zecilor sau sutelor, în cel mai rău caz.

Clasificatoarele Nearest Neighbour (NN) [1, p. 133-138] sunt cei mai folosiţi algoritmi ce separă spaţiul caracteristicilor în regiuni corespunzătoare fiecărei categorii. Ideea de bază este că documentele similare ar trebui să aparţină aceleiaşi clase. Faza de antrenare presupune o simplă indexare a datelor de test în funcţie de clasa din care fac parte – deci este foarte rapidă. Pentru clasificarea unui document dq, se determină k documente cât mai similare cu dq, unde k e o constantă. Clasa căreia îi aparţin cele mai multe dintre aceste documente reprezintă clasa în care este încadrat dq. O variantă mai bună este de a utiliza un scor pentru fiecare categorie astfel: dacă d face parte din cele k documente cele mai apropiate de documentul de clasificat dq şi d este clasificat în categoria cd, atunci scorul acestei categorii acumulează valoarea similarităţii dintre d şi dq. Clasa ce va avea scorul cel mai mare, după ce sunt considerate toate cele k documente cele mai apropiate de dq, va fi clasa în care acesta este încadrat. Pentru o acurateţe şi mai bună, pentru fiecare clasă se poate folosi un deplasament, offc, ce va fi adunat scorului categoriei. Considerând diverse valori pentru offc şi k, ce se obţin în urma antrenărilor succesive, clasificatorul poate deveni foarte puternic, oferind „o acurateţe comparabilă cu cea a celor mai bune clasificatoare de texte" [1, p. 135]. Avantajele clasificatoarelor NN sunt faptul că sunt uşor de antrenat şi de folosit, nefiind nevoie de construcţii matematice suplimentare. Pe lângă aceasta, oferă şi rezultate bune, dacă k şi deplasamentele categoriilor sunt bine alese. Dezavantajele lor ţin de timpii mari de clasificare, precum şi de cantitatea mare de memorie consumată în cadrul procesului de clasificare. Pentru a îmbunătăţi performanţele clasificatoarelor NN se poate folosi un algoritm greedy de selecţie a caracteristicilor esenţiale pentru fiecare clasă [1, p. 136-143].

Algoritmii ce folosesc estimarea unei distribuţii de probabilitate pentru fiecare clasă sunt puternici în clasificarea textelor, producând rezultate foarte bune. Dar pentru a obţine aceste rezultate este necesar un număr mare de date în setul de antrenare, pentru a obţine un estimator cât mai bun al fiecărei categorii. În condiţiile unui set de antrenament suficient de mare şi a unor categorii bine definite, categorizatoarele de texte bazate pe reţele Bayes şi SVM-uri au avut o acurateţe de peste 80% [3].

PORTAL WEB DE ŞTIRI INTELIGENT PENTRU LIMBA ROMÂNĂ

Prin combinarea facilităţilor oferite de tehnologiile web precum sindicalizarea şi a avantajelor furnizate de către tehnicile de minerit ale textelor este posibilă dezvoltarea unui portal de ştiri care poate funcţiona cu un minim de intervenţie umană. Acesta oferă o alternativă viabilă pentru portalurile de ştiri tradiţionale, datorită avantajelor oferite precum autonomia faţă de un administrator uman,

Figura 1 Grup de ştiri corespunzător unui subiect în cadrul portalului

precum şi metodologia folosită pentru a prezenta ştirile pornind de la importanţa unui subiect.

Funcţionarea şi arhitectura portalului

Una dintre trăsăturile fundamentale ale aplicaţiei este lipsa intervenţiei umane în timpul funcţionării – portalul trebuie să funcţioneze automat. Aceasta face ca portalul de ştiri construit să fie considerat un agent autonom.

Funcţionarea de bază a portalului web poate fi descrisă de către următoarea secvenţa de paşi:

Inspectarea periodică a siturilor agenţiilor de ştiri şi ziarelor ce oferă RSS şi colectarea automată a sindicărilor noi;

Introducerea ştirilor din cadrul fiecărui RSS nou într-o bază de date, pentru a oferi acces mai uşor la aceste informaţii;

Prelucrarea informaţiei text a fiecărei ştiri noi, prin aplicarea diverselor tehnici lingvistice, pentru determinarea vectorului caracteristic – numărul de apariţii ale fiecărui cuvânt;

Gruparea ştirilor folosind un algoritm de clustering pentru texte, pornind de la reprezentarea ştirilor în spaţiul m-dimensional al cuvintelor;

Clasificarea fiecărui grup de ştiri în cadrul unei categorii predefinite, folosind un clasificator care este reantrenat în mod regulat;

Generarea automată a paginilor web corespunzătoare subiectelor celor mai importante dintr-o anumită perioadă, grupate în diverse modalităţi, inclusiv pe fiecare categorie de ştiri. Aceste pagini web constituie rezultatul final al funcţionării portalului şi sunt vizibile de către utilizatori;

Generarea automată a sindicărilor RSS şi Atom oferite de portal, în conformitate cu subiectele folosite la pasul 6.

Acţiunile elementare descrise anterior, pot fi rulate atât într-o singură etapă, în mod secvenţial, urmând ordinea precizată, cât şi individual, la momente de timp diferite. Modul de funcţionare este determinat de cantitatea fluxurilor RSS preluate, de numărul ştirilor noi într-o anumită perioadă, cât şi de intervalul de timp la care situl este modificat automat.

Din modul de funcţionare al portalului descris anterior se poate remarca că acesta este un agent reactiv (răspunde schimbărilor din mediul său), necomunicativ (poate comunica numai într-o mică măsură, prin sindicalizare web) şi neadaptiv. Pe lângă acest comportament de agent autonom, portalul trebuie să implementeze şi un modul web care să permită căutarea text în cadrul subiectelor, cât şi o personalizare a portalului pe bază de utilizator, prin alegerea categoriilor, a ordinii acestora şi a numărului de ştiri din fiecare categorie.

Intervenţia unui administrator al portalului este necesară doar într-un singur punct şi anume pentru administrarea listei de adrese web a sindicărilor colectate în prima etapă a funcţionării automate. Legat de administrarea listei de sindicări folosite, administratorul va decide care dintre fluxurile RSS vor fi folosite pentru etapa de antrenare a clasificatorului de ştiri.

O etapă importantă în cadrul funcţionării aplicaţiei o constituie antrenarea iniţială şi, eventual, reantrenarea periodică, a clasificatorului. Antrenarea iniţială trebuie efectuată în momentul când avem exemple suficiente pentru fiecare categorie considerată. Reantrenarea se poate face periodic, pe măsură ce numărul setului de antrenament va creşte – în acest fel acurateţea clasificatorului ar trebui să se îmbunătăţească.

După cum se poate deduce şi din descrierea funcţionării portalului, acesta este alcătuit din două module cu o funcţionare relativ independentă: un modul ce conţine partea de agent a portalului (partea inteligentă) şi un modul web. Ambele module folosesc o bază de date în

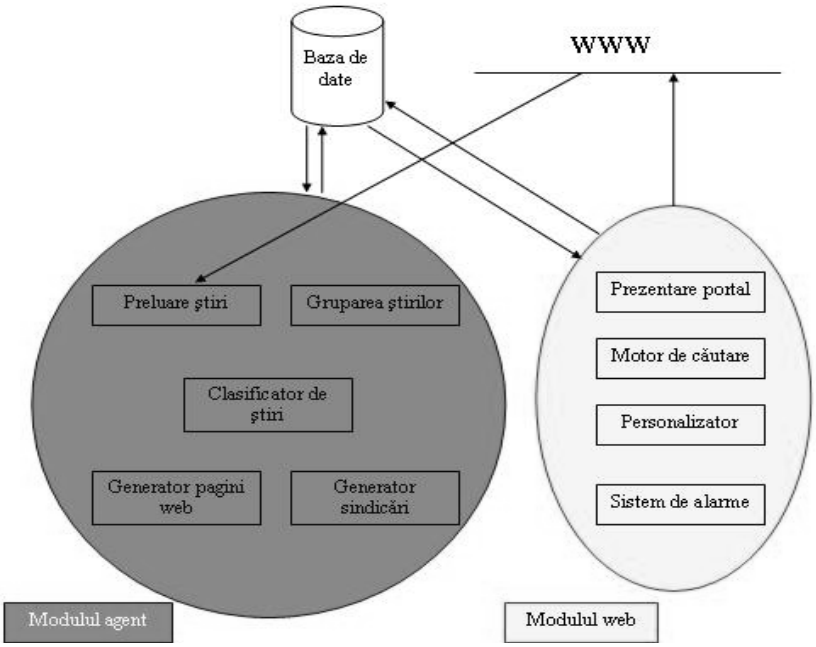

Figura 2 Arhitectura portalului web de ştiri – modulele agent şi web

care sunt salvate informaţiile despre ştiri, după cum se poate observa şi din figura 2.

Modulul agent este, la rândul său alcătuit din următoarele submodule:

Submodul de preluare a ştirilor – responsabil cu citirea din baza de date a listei de adrese web ale sindicărilor web, preluarea RSS-urilor de pe Internet şi salvarea acestora atât ca fişiere de sine stătătoare, pentru arhivă, cât şi în baza de date. Rezultă că acest submodul este responsabil şi de parsarea fişierelor RSS;

Submodul de grupare a ştirilor în subiecte – se ocupă cu citirea celor mai recente ştiri din baza de date, aplicând apoi un algoritm de grupare pentru determinarea subiectelor. Grupurile sunt apoi inserate în baza de date;

Clasificatorul de ştiri – funcţionează în două etape: prima etapă o constituie antrenarea folosind un set de date preluate din RSS-uri predefinite pentru fiecare categorie, iar a doua etapă constă în clasificarea subiectelor determinate la pasul anterior;

Generatorul de pagini web – creează paginile web cu ştirile cele mai importante din fiecare categorie, paginile corespunzătoare fiecărui subiect, precum şi alte pagini web.

Generatorul de sindicări – rulează odată cu submodulul anterior, şi creează fişierele RSS şi Atom folosite pentru sindicalizarea conţinutului portalului.

Modulul web reprezintă interfaţa dintre portal şi utilizatori şi conţine patru submodule:

Interfaţa (prezentarea) portalului – conţine paginile web şi sindicările generate de către submodulele 4 şi 5 ale modulului agent, împreună cu alte pagini web statice, imagini şi fişiere necesare funcţionării portalului web;

Motor de căutare – permite căutarea în cadrul subiectelor şi ştirilor a unor cuvinte cheie;

Submodul de personalizare – permite personalizarea paginii de start a portalului în funcţie de interesele utilizatorului;

Sistem de alarme – definirea de cuvinte cheie pentru trimiterea de alarme prin poştă electronică la întâlnirea acelor cuvinte în cadrul ştirilor.

Cele două module pot funcţiona independent, fiecare pe câte un server. De asemenea, baza de date poate să fie găzduită pe un server de sine stătător. De altfel, datorită volumului mare de date pe care portalul le va aduna şi prelucra de-a lungul funcţionării, o astfel de arhitectură este foarte utilă, chiar necesară. În plus, arhitectura are un grad mare de paralelism, întrucât modulul agent poate să fie la rândul său distribuit pe mai multe calculatoare. Submodulele constituente pot să funcţioneze independent, necesitând doar acces la baza de date.

Prelucrarea textului şi gruparea ştirilor

Fluxurile de ştiri trebuie sa fie procesate înainte de aplicarea metodelor de grupare şi clasificare. Deoarece diacriticele nu sunt folosite de către toate sursele de ştiri, acestea au trebuit sa fie eliminate, indiferent de schema de codificare utilizată de furnizor. De asemenea, sunt eliminate etichetele şi entităţile HTML, precum şi cuvintele de stop. Textul rezultat este descompus în cuvinte şi fiecare termen este adus la forma de *stem*, folosind un algoritm special pentru limba română, pentru a reduce numărul de forme ale cuvintelor. Implementarea unui algoritm de eliminare a sufixelor pentru limba română este extrem de dificilă deoarece regulile de flexionare sunt numeroase şi destul de complicate, afectând şi structura internă a cuvintelor. Mai mult, fiecare regulă de eliminare a sufixelor poate aduce dezavantaje [14], motiv pentru care s-a optat pentru folosirea unui set restrâns de reguli. Acestea reduc numărul de termeni cu aproximativ 20-25%, depinzând de numărul de ştiri procesate.

	Spaţiu frecvenţă			Spaţiu boolean		
Nr. ştiri	Nr. grup-uri	Dim. medie	Dim. max.	Nr. grup-uri	Dim. medie	Dim. max.
666	545	1.22	10	556	1.20	6
674	524	1.29	14	510	1.32	38
641	530	1.21	12	532	1.20	13
545	440	1.24	12	430	1.27	35
644	520	1.24	22	533	1.21	15
780	650	1.20	14	641	1.22	39
1024	828	1.22	18	845	1.20	10

Tabelul 1 Rezultatele grupării

După etapa de procesare iniţială prezentată anterior, se extrage vectorul caracteristic pentru fiecare articol. Deoarece spaţiul caracteristicilor este foarte mare, iar o ştire conţine doar un număr mic de termeni, aceşti vectori sunt foarte rari, motiv pentru care este utilă o reprezentare folosind un set ordonat după termeni.

Algoritmul de grupare foloseşte o strategie ierarhică aglomerativă, cu asignare tare şi cu un calcul de tip legătură medie pentru similaritatea între grupuri. La fiecare pas, două grupuri sunt unite numai dacă similaritatea lor este mai mare decât un prag determinat în faza de testare a funcţionării portalului. De fapt, algoritmul utilizează două praguri diferite, o valoare mai mare pentru creearea unui prim set de grupuri cu articole foarte similare şi o valoare mai scăzută ce este utilizată numai când nu se mai găsesc grupuri care să aibă o similaritate mai mare decât prima valoare. Pragul mai mare evită unificarea, într-o primă fază, a articolelor care au subiecte asemănătoare, dar totuşi diferite. În acest fel, fiecare grup format în prima fază va acoperi doar un singur subiect. Pentru îmbunătăţirea timpului de procesare, a fost folosit un *buffer* pentru memorarea similarităţilor între grupuri.

Deoarece modalitatea de calcul a similarităţii este foarte importantă pentru gruparea ierarhică, au fost evaluate rezultatele folosind toate măsurile prezentate în [10], pentru a le găsi pe cele mai potrivite. Pentru spaţiile vectoriale ale caracteristicilor bazate pe frecvenţă, a fost aleasă similaritatea cosinus, iar pentru cele booleene a fost aleasă o măsură definită de către autori şi adaptată din teorema cosinusului. Rezultatele obţinute folosind aceste metode sunt prezentate în tabelul 1.

Clasificarea subiectelor de ştiri

Determinarea celor mai importante subiecte ale zilei îmbunătăţeşte calitatea informaţiei oferite consumatorului de ştiri. Dar sunt foarte mulţi utilizatori care doresc să citească numai ştirile dintr-un anumit domeniu de interes. Pentru aceasta, a fost implementat un clasificator automat pentru grupurile de ştiri, având următoarele opt categorii: *România*, *Politică*, *Economie*, *Cultură*, *Internaţional*, *Sport*, *High Tech* şi *Monden*. Criteriile determinante în alegerea categoriilor au fost: categoriile să fie cât mai bine definite (să nu se suprapună), să includă domeniile generale de interes şi să existe fluxuri de ştiri specializate pentru fiecare categorie, care să poată fi utilizate în cadrul antrenării clasificatorului.

Trei variaţii ale clasificatoarelor de tip Nearest Neighbour au fost implementate şi comparate: prima este un k-NN simplu, iar cea de a doua foloseşte un clasificator k-NN cu un scor pentru fiecare categorie, ce este calculat prin însumarea similarităţilor celor mai apropiate k documente ce aparţin aceleiaşi clase. A treia metodă seamănă cu un clasificator de tip NN, dar foloseşte o abordare un pic diferită. În loc de a calcula similaritatea între documentul ce trebuie să fie clasificat şi cei mai apropiaţi vecini ai săi, algoritmul calculează similaritatea între acest document şi centroidul fiecărei categorii, alegând categoria de al cărui centroid este cel mai similar. Din această cauză, metoda aceasta se numeşte cel mai apropiat centru (*nearest centre – NC*) sau NN bazat pe centroid şi funcţionează foarte bine când obiectele dintr-o categorie nu sunt distribuite uniform în jurul centroidului său. Spre deosebire de clasificatoarele NN clasice, această metodă necesită o fază de antrenare simplă ce calculează vectorul caracteristic al fiecărei categorii prin însumarea vectorilor ce aparţin clasei respective din setul de antrenare.

Categorie	Dimensiune set de antrenare	Dimensiune set de validare
Romania	640	325
Politică	313	157
Economie	182	92
Cultură	88	45
Internaţional	481	241
Sport	316	158
High-Tech	76	38
Moden (HL)	84	43
Total	2180	1099

Tabelul 2 Dimensiunile seturilor de antrenare şi de evaluare folosite pentru evaluarea clasificatoarelor

Deşi în alegerea categoriilor s-a încercat să se profite de existenţa unor fluxuri de ştiri specifice categoriilor, acest lucru nu a fost posibil pentru fiecare categorie, motiv pentru care a fost necesară o mapare a categoriilor furnizorilor de ştiri către cele ale portalului. În acest fel, a fost posibilă construirea semi-automată a datelor din setul de antrenare. Clasificatoarele au fost comparate folosind tehnica de validare încrucişată, pentru a determina metoda cu cea mai bună acurateţe. Astfel, datele din setul de antrenament, au fost împărţite în două subseturi: două treimi au fost folosite pentru antrenarea efectivă, iar ultima treime a fost utilizată pentru evaluarea rezultatelor. Dimensiunile fiecarui subset, pentru fiecare categorie, sunt prezentate în tabelul 2.

Numărul de articole din fiecare categorie nu este foarte mare şi acestea sunt distribuite inegal, existând o diferenţă mare între categoria cu cele mai puţine articole în setul de antrenament şi cea cu cele mai multe. Acestea au afectat procesul de antrenare, rezultatul reflectându-se astfel în performanţa clasificatoarelor. Parametrii prezentaţi în continuare se pot îmbunătăţi odată cu creşterea dimensiunii corpusului de antrenament.

Utilizând datele menţionate mai sus, au fost testate următoarele clasificatoare: unul de tipul cel mai apropiat centru (NC), unul de tipul NN şi două de tipul k-NN (pentru k = 3 şi k = 5). Parametrii evaluaţi au fost: acurateţea clasificatorului definită ca numărul de articole clasificate corect împărţit la numărul total de ştiri, durata procesului de antrenare şi durata procesului de clasificare efectivă. În plus, pentru fiecare clasificator a fost calculată matricea de confuzie, dar aceste date nu vor fi prezentate în această lucrare. Pentru fiecare clasificator, au fost utilizate ambele tipuri de criterii de similaritate.

După cum se poate observa şi din tabelul 3, ce prezintă acurateţea clasificatoarelor, modelele bazate pe frecvenţă sunt mai precise decât cele binare, indiferent de algoritmul folosit. Cea mai bună acurateţe este oferită de către algoritmul cel mai apropiat centru (NC), care depăşeşte algoritmii NN şi k-NN, chiar şi variantele care folosesc însumarea similarităţii celor mai apropiaţi vecini din fiecare categorie. Deşi faza de antrenare a algoritmului NC necesită mai mult timp decât cea a algoritmilor NN şi k-NN, acesta compensează prin viteză de clasificare, fiind de câteva ori mai rapid decât algoritmii de tip NN.

Tipul clasificatorului	Timp antrenare (s)	Timp clasificare (s)	Acurateţe (%)
NC – cosinus	279	17	64
NC – binar	280	18	63
NN – cosinus	1	60	61
3-NN – cosinus	1	59	57
5-NN – cosinus	1	59	54
3-NN – cosinus (sumă)	1	59	58
5-NN – cosinus (sumă)	1	59	56

Tabelul 3 Acurateţea clasificatoarelor implementate

Clasificatorul de tip NC, folosind vectorul caracteristic al frecvenţelor şi similaritatea cosinus, s-a dovedit cea mai bună opţiune atât din punctul de vedere al acurateţii, cât şi a duratei clasificării. În figura 3 este prezentată matricea de confuzie a acestui clasificator – ultima coloană conţine precizia, iar ultima linie acoperirea, pentru fiecare categorie. În plus, parametrii săi descriptivi caracteristici sunt următorii: acoperirea medie = 59%, precizia medie = 62%, acurateţea = 64%, F1 (media armonică a preciziei şi acoperirii) = 61%.

	Rom	Pol	Eco	Cul	Int	Spo	Teh	HL	**Tot**	Prec
Rom	177	28	32	11	37	12	10	18	**325**	0.54
Pol	15	117	10	0	12	2	1	0	**157**	0.75
Eco	16	2	57	2	10	1	2	1	**92**	0.63
Cul	11	1	0	21	3	3	2	3	**45**	0.48
Int	37	12	16	4	162	3	3	4	**241**	0.67
Spo	5	3	7	4	7	128	1	3	**158**	0.81
Teh	1	0	9	1	0	0	27	0	**38**	0.71
HL	6	0	5	10	4	0	2	16	**43**	0.37
Tot	**268**	**163**	**136**	**53**	**235**	**149**	**48**	**45**		
Rec	0.66	0.71	0.42	0.40	0.69	0.86	0.56	0.36		

Figura 3 Matricea confuziei pentru clasificatorul de tip NC

CONCLUZII

Lucrarea prezintă o alternativă pentru portalurile de ştiri clasice, proiectată pentru a rezolva problemele cantităţii mari de ştiri şi a redundanţei informaţiei, prin folosirea acesteia din urmă ca un avantaj. În plus, portalul foloseşte sindicalizarea web şi tehnici de prelucrare a limbajului natural pentru a atinge o funcţionare autonomă, independentă de intervenţia umană, ce poate fi numită inteligentă deoarece determină automat importanţa unui subiect, precum şi categoria sa.

Fluxurile de ştiri oferă o soluţie simplă pentru preluarea ştirilor de la un număr mare de surse şi gruparea este folosită pentru a exploata ştirile similare şi pentru a le grupa în funcţie de subiect. Utilizatorului îi sunt prezentate subiectele de ştiri, ordonate în funcţie de numărul de ştiri ce alcătuiesc fiecare subiect. În plus, cititorii au posibilitatea de a alege sursa favorită pentru a citi un articol dintr-un anumit subiect.

Clasificarea automată a subiectelor de ştiri aduce avantaje importante faţă de clasificarea unei singure ştiri deoarece caracteristicile unui grup sunt mai consistente decât cele ale unui singur articol. Acest lucru este de o importanţă crucială în aplicaţie deoarece sindicalizarea oferă numai o scurtă descriere pentru fiecare ştire în parte.

Dezvoltarea ulterioară a portalului urmăreşte două direcţii esenţiale: îmbunătăţirea tehnicilor de grupare şi de clasificare, precum şi încercarea de a face funcţionarea sa independentă de limbă sau, cel puţin, multilingvă.

PRECIZĂRI

Cercetările prezentate în această lucrare au fost finanţate parţial din proiectul european FP7 STREP LTfLL şi din proiectul naţional CNCSIS K-Teams.

REFERINŢE

1. Chakrabarti, S.: Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers (2002)

2. del Corso, G., Gulli, A., Romani, F.: Ranking a Stream of News. Proceedings of the 14th international conference on World Wide Web, Chiba (2005) 97–106

3. Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization. Proceedings of the Seventh International Conference on Information and Knowledge Management, Bethesda (1998) 148–155

4.

5.

6. Gabrilovich, E., Dumais, S., Horvitz, E.: Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty. Proceedings of the Thirteenth International World Wide Web Conference (2004)

7. Internet World Stats: World Internet Usage and Population Statistics, disponibil online la adresa http://www.internetworldstats.com/stats.htm (2007)

8. Manning, C., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, Massachusetts (2003)

9. Netcraft: February 2008 Web Server Survey, disponibil online la adresa http://news.netcraft.com/archives/web_server_survey.html (2008)

10. Page, L., Brin, S., Motwani, R. and Winograd, T.: The page-rank citation ranking: Bring-ing order to the web. Technical Report, Stanford University (1998)

11. Porter, M.F.: An algorithm for suffix stripping. Journal of the Society for Information Science, 3(14) (1980) 130–137

12. Strehl, A.: Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining. Doctoral dissertation, University of Texas at Austin (2002)

13. Toda, H., Kataoka, R.: A Clustering Method for News Articles Retrieval System. Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba (2005)

14. Ueda, Y., Oka, M., Yamashita, A.: Evaluation of the Document Categorization in "Fixed-point Observatory". Proceedings of NTCIR-5 Workshop Meeting, Tokyo (2005)

15. Ullman, J.: Data Mining Lecture Notes (2000)

16. Yahoo! Search Blog: Our Blog is Growing Up And So Has Our Index, disponibil online la adresa http://www.ysearchblog.com/archives/000172.html (2005)