Revista Română de Interacţiune Om-Calculator 7 (3) 2014, 209-224

Modelarea, analiza şi clasificarea conversaţiilor în medii colaborative

Alexandru Bardaş[1], Mihai Dascălu[1], Ştefan Trăuşan-Matu[1,2]

[1]Universitatea Politehnica din Bucureşti, Facultatea de Automatică şi Calculatoare
Bd. Splaiul Independenţei, Nr. 313, 060042, Bucureşti, România

[2]Institutul de Cercetări în Inteligenţă Artificială
Calea 13 Septembrie, Nr. 13, 050711, Bucureşti, România
E-mail: alex.bardas@gmail.com, mihai.dascalu@cs.pub.ro, stefan.trausan@cs.pub.ro

Rezumat. Clasificarea conversaţiilor în medii colaborative este necesară pentru o mai bună înţelegere a subiectelor discutate. Ontologiile reprezintă o modalitate eficientă şi reprezentativă de conceptualizare a unui domeniu. Plecând de la un set predefinit de ontologii specifice mai multor domenii, se doreşte clasificarea cât mai precisă a unei conversaţii purtate între mai mulţi participanţi. Pentru a facilita acest proces, este nevoie ca limba în care mesajele sunt scrise să fie detectată şi intervenţiile utilizatorilor să nu conţină greşeli gramaticale. Astfel, termenii utilizaţi de către participanţii la conversaţii sunt regăsiţi în conceptele ce definesc ontologiile, iar conceptele predominante şi centrale determină domeniile pe baza cărora conversaţiile sunt clasificate. Pentru a realiza această analiză, am construit o platformă web colaborativă ce permite schimbul de mesaje între utilizatori în timp real, vizualizarea termenilor regăsiţi în ontologii cu ajutorul unui graf radial şi exportul arhivei de discuţii într-un format uşor accesibil care să faciliteze clasificarea lor ulterioară. Aplicaţia este construită pentru a permite accesul de pe orice tip de dispozitiv, afişând informaţiile minime necesare pentru un mobil şi redimensionându-se automat pentru un ecran mai mare. În articol sunt descrise implementarea aplicaţiei, avantajele folosirii acesteia şi ale clasificării automate, precum şi oportunitatea integrării ei în diverse contexte educaţionale.

Cuvinte cheie: mesagerie instantanee, conversaţie, colaborare, clasificare, corecţie ortografică, ontologie.

1. Introducere

Web-ul social este centrat pe interactivitatea dintre utilizatori, precum şi pe activităţi de publicare, căutare, regăsire şi interpretare a informaţiilor. Implementarea unei infrastructuri care să permită colaborarea cât mai eficientă şi mai rapidă sau care să facă atât posibilă extragerea de date cu privire la profilele şi schemele de comunicare ale utilizatorilor, cât şi

prelucrarea acestora, a devenit o necesitate în cazul marilor comunități online. Scopul acestor facilități este de a-şi ajuta utilizatorii să transmită informații relevante, să cunoască alte persoane cu aceleaşi interese sau să găsească soluții la problemele pe care le întâmpină.

Deşi world wide web-ul a fost proiectat inițial ca un mediu pur informațional, acesta a evoluat spre un mediu al aplicațiilor, marcând tranziția către un web semantic (Berners-Lee et al., 2001). În prezent, aceste aplicații reprezintă sisteme software complexe care oferă servicii interactive şi personalizabile.

Companii cu renume investesc sume considerabile în atragerea cu orice preț a unui număr cât mai mare de utilizatori şi oferirea posibilității de colaborare între aceştia. În noiembrie 2011 Microsoft a achiziționat Skype pentru 8.5 miliarde de dolari (Bright, 2011), iar în februarie 2014, Facebook a achiziționat aplicația de mesagerie instantanee WhatsApp pentru o tranzacție estimată la peste 19 miliarde dolari (Stone, 2014). Aceste achiziții ilustrează cerința foarte mare care există pentru comunicații în timp real şi retenția utilizatorilor. Astfel, relaționând datele deja existente sau înțelegând contextul conversațiilor purtate între utilizatori, precum şi graful rețelei de prieteni, domeniile de interes ale acestora pot fi identificate mult mai bine.

Informațiile sunt atât de numeroase încât prelucrarea sau clasificarea lor nu mai poate fi făcută manual, de către oameni. Conform McAfee & Brynjolfsson (2012), în 2012 aproximativ 2.5 exabytes de date erau creați în fiecare zi, iar volumul era estimat că se dublează la fiecare 40 luni. Suplimentar, conform Cisco (2013), în 2017 peste 3.6 miliarde de utilizatori (mai bine de 48% din populația preconizată la acel moment) vor fi conectați la Internet. Vor fi aproximativ 1.4 dispozitive mobile pe cap de locuitor, fapt ce reliefează o dependență din ce în ce mai mare vizavi de tehnologie.

O adopție mai mare va determina şi o creştere în ceea ce priveşte cantitatea de informație disponibilă pe Internet. Este însă interesant de observat ce se va întâmpla cu toate datele noi care apar, deoarece şi interesul pentru gruparea şi înțelegerea semantică a documentelor a crescut foarte mult. Mai mult, majoritatea motoarelor de căutare indexează în continuare documentele în funcție de conținutul lexical, şi doar parțial în funcție de semantica aferentă.

Astfel, trebuie găsită o alternativă ca aceste date să fie prezentate atât într-un format accesibil utilizatorilor, cât şi sistemelor de calcul ce urmează să le prelucreze şi să le clasifice. Din acest motiv, definirea lor semantică

devine foarte importantă. Aceasta va facilita managementul cunoştinţelor pe web şi în particular, conectarea şi reutilizarea cunoaşterii (sindicalizarea conținutului), dar şi localizarea cunoştinţelor noi şi relevante, cu ajutorul sistemelor de recomandare.

Lucrarea de față realizează o trecere în revistă a aplicațiilor colaborative existente, detaliază arhitectura soluției propuse şi facilitățile integrate de prelucrare a limbajului natural, precum şi aplicabilitatea aplicației dezvoltate în diverse scenarii educaționale.

2. Platforme web colaborative

În general, colaborarea facilitează munca în echipa şi oferă posibilitatea creşterii productivității, alocării mai bune a resurselor şi dobândirii mai uşoare de informații noi (Stahl, 2006). O aplicație colaborativă poate fi văzută ca un program accesat din Internet sau dintr-o rețea, care oferă o serie de funcționalități menite să ajute la dezvoltarea şi îndeplinirea rapidă a sarcinilor de lucru.

Principalele categorii de software colaborativ sunt destinate:

- coordonării: calendare, foi de calcul online, sisteme de management al proiectelor;
- comunicării: wiki, email, blog;
- consfătuirii: mesageria instantanee, forum.

Pentru a argumenta utilizarea unei anumite tehnologii în cadrul experimentelor educaționale derulate în cadrul facultății, s-a considerat necesară analiza minuțioasă a celor mai bune alternative web disponibile. Multe platforme online au rolul de a pune în legătură utilizatorii prin oferirea unor soluții eficiente, cât mai uşor de utilizat, furnizând totodată seturi diferite de funcționalități. Modalitățile de colaborare sunt diferite, de la soluții de tipul întrebi şi primeşti răspuns (Stackoverflow, Quora), la aplicații de management de proiect (Asana, Jira, Trello), până la soluții ce oferă mesagerie instantanee (Campfire, HipChat, Slack - Figura 1 şi Appear.in - Figura 2).

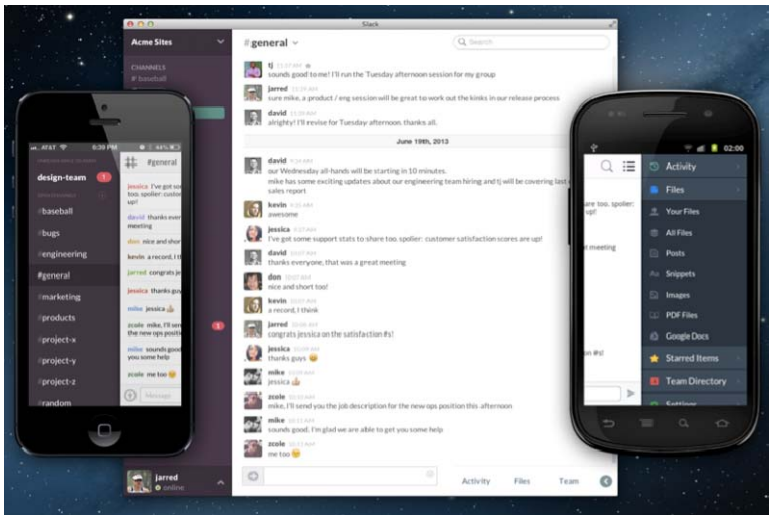


Figura 1. Interfața aplicației Slack reliefând portabilitatea platformei

O soluție populară este reprezentată de chatul prin canalale IRC care reprezintă o alternativă matură, testată în timp, simplă şi eficientă. Totodată, pe măsura trecerii timpului şi o dată cu apariția alternativelor, limitările au apărut şi s-au accentuat tot mai clar. Multe canale nu sunt arhivate, discuțiile purtate fiind practic pierdute. De asemenea, căutarea unor topice sau termeni poate fi uneori dificilă, alteori imposibilă. De cele mai multe ori, domeniul de discuție al camerei este greu de identificat. Comunicarea prin Internet Relay Chat (IRC – RFC 1459) este o soluție depăşită pentru cei care vor să analizeze subiectele discutate sau să colaboreze cât mai eficient.

Slack este o soluție relativ nouă de mesagerie instantanee, cu o adopție foarte bună încă de la început, de la creatorul platformei Flickr. A apărut din nevoia de a oferi ceva nou față de aplicațiile disponibile pe piață (ex: Campfire, Hipchat). Este rapid şi foarte uşor de utilizat, oferind o interfață simplistă şi intuitivă. Este destinat exclusiv colaborării într-o organizație, pentru un grup ridicat de persoane. Oferă căutare de termeni atât în cadrul discuțiilor, dar şi al documentelor încărcate. Căutarea unui topic se efectuează uşor, putând, de exemplu, căuta mai întâi dacă au mai fost semnalate probleme asemănătoare înainte de a le mai adresa încă o dată. Este disponibil atât în versiune web, dar şi ca aplicație nativă pe toate sistemele de operare. Succesul care îl are dovedeşte că multe companii sunt

dispuse să plătească pentru o aplicație destinată atât comunicării, cât şi organizării discuțiilor unui grup.

Appear.in este o platformă de ultimă generație care foloseşte cele mai noi tehnologii şi API-uri web (WebRTC) pentru a oferi pe lângă funcționalitatea de chat, şi facilități video şi audio. O cameră poate găzdui o conferință video cu până la 8 participanți. Un mare avantaj pe care îl oferă este că înregistrarea pe sit este opțională, oricine putând accesa o cameră de discuții doar ştiind numele acesteia. Soluția este compatibilă doar cu cele mai noi versiuni ale browserelor web şi demonstrează faptul că web-ul este un mediu matur şi suficient de puternic pentru a oferi funcționalitatea şi performanța unui mediu nativ.

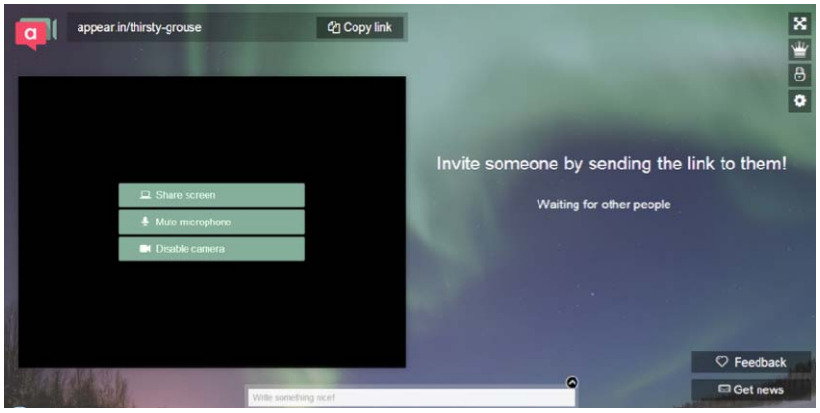


Figura 2. Interfața aplicației Appear.in

3. Facilități integrate de prelucrare a limbajului natural

În vederea clasificării discuțiilor, detecția limbii este un prim pas important. Cu ajutorul acesteia, se pot oferi sugestii adecvate de corecție şi se pot căuta diverşi termeni în ontologiile specifice anumitor domenii. Astfel, conversațiile purtate într-un mediu colaborativ pot fi clasificate în funcție de domeniul de discuție. Există mai mulți algoritmi folosiți pentru această clasificare, cei mai mulți fiind bazați pe metode statistice. Detalierea acestora se regăseşte în secțiunea 3.1.

Totodată, întrucât în cadrul unei conversații pot exista vorbitori non-nativi ai limbii utilizate, am optat pentru integrarea unui sistem de traducere, din limba detectată într-o alta limbă selectată de către utilizator. S-a

constatat că cele mai bune astfel de motoare sunt disponibile ca servicii web (Google Translate, Bing Translator, Babel Fish). Implementarea în aplicație a unui algoritm eficient nu este un aspect trivial, iar rezultatele ar fi fost sub cele ale serviciilor oferite de cele web menționate. În final, am decis integrarea serviciului *Bing Translator,* ce oferă suport pentru 45 limbi. Spre deosebire de alternative, *Bing Translator* este disponibil gratuit şi pune la dispoziție o documentație amplă, aspecte ce au favorizat integrarea acestuia.

Ulterior, procesul de corecție ortografică este aplicat înainte de a realiza o clasificare automată a discuțiilor purtate. Ambele facilități sunt descrise în detaliu în sub-secțiunile următoare.

3.1 Detecția limbii bazată pe modelul n-gram

Una dintre cele mai eficiente metode statistice de detectare a limbii presupune crearea unui model n-gram (secvențe de *n* caractere). Metoda descrisă de Cavnar & Trenkle (1994) se bazează pe faptul că pentru fiecare limbă există un set de cuvinte dominante din prisma frecvenței ridicate de apariție a lor în cadrul textelor. Astfel, textele vor avea distribuții ale frecvenței n-gram asemănătoare dacă sunt din acelaşi domeniu şi limbă.

Modelul inițial este construit dintr-un text de învățare specific fiecărui limbaj. Pentru un text nou se creează un alt model, iar acesta este comparat cu fiecare model stocat anterior (aferent fiecări limbi). Limba cea mai probabilă este cea care are modelul cel mai asemănător cu modelul creat pentru noul text. Cu cât textul este mai lung, cu atât detecția este mai precisă. Problemele apar mai ales în cazul textelor scurte sau când sunt folosite mai multe limbi în acelaşi text, caz în care algoritmul poate fi ineficient.

Algoritmul funcționează rapid şi în cele mai multe cazuri (mesaje corecte, mai lungi de 100 caractere) oferă rezultate relevante. Astfel, acuratețea metodei este foarte bună, detecția având o rată de succes între 96 şi 100% pentru diverse texte şi limbi.

3.2 Corecție ortografică

Pentru detectarea greşelilor de ortografie şi pentru oferirea de sugestii este folosit un algoritm statistic asemănător cu cel descris de Norvig (2014). Pentru a antrena acest algoritm este nevoie de un set destul de mare de texte şi de dicționarul specific unei limbi. Detecția limbii unui mesaj efectuată în

prealabil este necesară pentru a selecta setul de date corespunzător corecției ortografice. Astfel, pornind de la un corpus de texte, sunt extrase toate cuvintele, transformate în varianta cu minuscule şi este antrenat un model de probabilitate prin numărarea frecvenței de apariție a fiecărui cuvânt. Distanța de editare a 2 cuvinte sau distanța Levenshein (1966) este definită ca numărul de editări ce trebuie făcut într-un cuvânt pentru a-l transforma în cel de-al doilea cuvânt. Pentru un cuvânt *w*, sunt enumerate toate corecțiile posibile *c* care poate reprezenta un set foarte mare de cuvinte, iar apoi sunt păstrate doar corecțiile valide, care sunt regăsite în dicționarul limbii. Se consideră că 80-90% dintre greşelile de ortografie diferă de cuvântul corect doar printr-o literă.

Astfel, fiecare cuvânt care nu este găsit direct în dicționar va fi supus acestui algoritm pentru o distanță de editare egală cu 1. Dacă nu este găsită nicio corecție, va fi aplicat algoritmul şi pentru o distanță de editare egală cu 2, iar corecțiile întoarse sunt oferite ca sugestii de editare. Totodată, specific conversațiilor chat sunt abrevierile (spre ex., ”brb”, ”lol”) care sunt înlocuite automat pe baza unor liste predefinite de mapări.

3.3 Clasificarea discuțiilor

Fiecare conversație purtată într-o cameră virtuală de discuții este analizată ulterior pentru a fi clasificată într-un domeniu de interes. Algoritmul de clasificare este bazat pe frecvența aparițiilor cuvintelor cheie din arhiva inițială de discuții pe baza unui set de ontologii predefinite. Cu cât fiecare set este mai mare, cu atât creşte probabilitatea regăsirii a cât mai multor cuvinte în ontologii. Analiza se bazează pe faptul că atât mesajele, cât şi ontologiile sunt scrise şi definite în aceeaşi limbă.

Algoritmul funcționează astfel: pentru o arhivă A, se cunosc toate mesajele $m[1:N]$, iar pentru un mesaj m_i se ştie autorul m_{ia}, limba mesajului m_{il} şi data trimiterii lui m_{id}. Se găseşte limba predominantă L din lista de mesaje, iar replicile scrise în altă limbă sunt traduse în L (folosind serviciul Bing Translator). Pentru o listă inițială de ontologii $O[1:M]$ definite în L, se creează 2 dicționare: (1) H_s, unde cheile sunt conceptele prezente în ontologii, iar valorile sunt liste cu ontologiile în care apar şi nivelul ierarhic pe care apar (2 ontologii diferite pot conține același concept) şi (2) H_o, unde cheile sunt domeniile ontologiilor, iar valorile vor fi setate ulterior.

Dacă un cuvânt w din m_i nu este găsit în dicționarul limbii L, atunci se aplică algoritmul de corecție cu o valoare a distanței de editare egală cu 1 şi

rezultatele sunt căutate în H_s. Dacă sunt găsite, pentru fiecare valoare v de la cheia din H_s găsită, $H_o[v]$ va fi incrementat cu o valoare $v_2 \cdot i_v$. Dacă w este găsit în L, este căutat în H_s şi dacă este găsit, pentru fiecare valoare de la cheia din H_s găsită, $H_o[v]$ va fi incrementat cu o valoare $v_1 \cdot i_v$, unde i_v este determinat în funcție de nivelul ierarhic în care termenul v apare în ontologie. Astfel, dacă se află pe nivelul n, i_v va fi $1/n$. Cu alte cuvinte, cu cât conceptul este mai apropiat de rădăcina ontologiei, cu atât acesta este mai general şi relevanța vizavi de un domeniul specific este diminuată (Resnik, 1995).

Algoritmul a fost testat pe diverse seturi de date constând în arhive de conversații cu sute/mii de mesaje şi ontologii definite în limba engleză. Pentru stabilirea valorilor optime v_1, v_2 s-a rulat algoritmul cu valori incrementale ale variabilelor. Astfel, s-au obținut rezultatele descrise în Tabelul 2, din care se observă că termenii regăsiți direct în ontologii sunt mai relevanți decât cei găsiți prin aplicarea corecției ortografice. Ontologiile adăugate în sistem sunt preluate dintr-o sursă publică, ce poate fi regăsită la http://semanticweb.org/wiki/Ontology şi descriu numeroase concepte şi domenii de activitate (ex: muzică, sporturi, date biografice, filme, etc). Totodată, în cadrul aplicației a fost integrată şi ontologia de Human-Computer Interaction dezvoltată în cadrul cursului de Interacțiune Om-Calculator din cadrul Universității Politehnica din Bucureşti, Departamentul Calculatoare.

Tabelul 2. Rezultatele experimentale pentru clasificarea mesajelor

Mesaje în arhivă	**Concepte în ontologii**	v_1	v_2	**Clasificare adecvată a arhivei într-un domeniu**
< 200	< 3000	0.25	*0.5*	Nu
< 200	< 3000	*0.5*	*0.75*	Nu
< 200	> 3000	*1*	*0.5*	Da
< 200	> 3000	*1*	*0.75*	Da
> 200	> 3000	*0.25*	*0.75*	Nu
> 200	> 3000	*0.75*	*0.25*	Da
> 200	> 3000	*0.75*	*0.5*	Da
> 200	> 3000	*1*	*0.5*	Da

În final, se aleg valorile maxime din $H_{o,}$ iar cheile corespunzătoare acestor valori sunt domeniile care au apărut cel mai frecvent în *A*. În cazul în care valorile sunt sub o anumită limită sau egale toate cu 0, se consideră că arhiva nu a putut fi clasificată.

4. Descrierea aplicației dezvoltate

Aplicația web dezvoltată facilitează comunicarea între doi sau mai mulți participanți şi este bazată pe camere de discuție, cu posibilitatea creării unui număr nelimitat de astfel de camere, dar şi a invitării de prieteni. Utilizatorii se pot înregistra în aplicație pentru a lua parte la discuțiile la care au acces sau pot inița sesiuni în camere noi. Astfel se pot crea 2 tipuri de canale de discuție: (1) *chat*, având scopul de a facilita comunicarea între 2 utilizatori (1:1), respectiv (2) *chatroom* pentru comunicarea între mai mulți utilizatori (1:N). Fiecare cameră are propriul istoric al discuțiilor, iar acestea se pot exporta oricând în format XML sau JSON, pentru a putea fi prelucrate şi analizate independent de aplicația curentă.

Spre deosebire de marea majoritate a aplicaților care oferă astfel de soluții, o funcționalitate aparte o reprezintă posibilitatea adăugării de legături explicite, facilitate regăsită şi în ConcertChat. În acest mod, orice mesaj dintr-o cameră poate fi referențiat în momentul adăugarii unei noi replici, iar toate referințele unui mesaj pot fi vizualizate în interfață, fiind marcate corespunzător. Un utilizator îşi poate edita un mesaj imediat dupa postare, putând aplica astfel corecții ortografice sau putând aduce o îmbunătățire semantică.

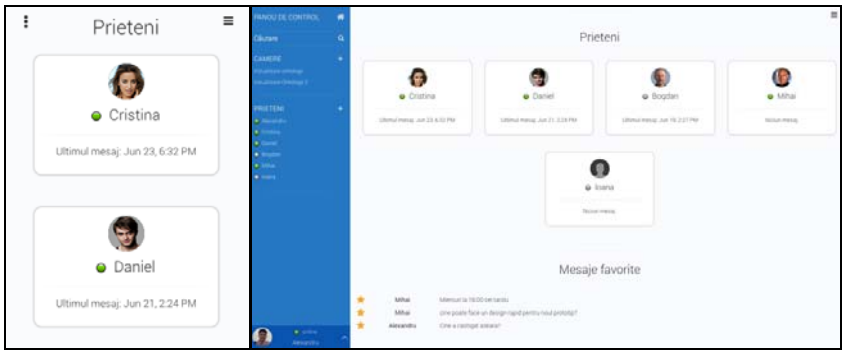
Interfața şi serverul au fost dezvoltate folosind JavaScript. Întrucât experiența oferită utilizatorului, precum şi viteza de încărcare şi navigare sunt criterii care au fost considerate necesare, s-a optat pentru dezvoltarea unei aplicații care încarcă o singură pagină la început, apoi îi modifică dinamic conținutul în funcție de interacțiunea utilizatorului cu aceasta (Single Page Application) prin intermediul websockets. Serverul pune la dispoziție un API (interfață destinată programatorilor) securizat, bazat pe permisiunile utilizatorilor şi acționează ca un proxy între interfață şi baza de date. Beneficiind de o astfel de arhitectură, aplicația poate fi transferată şi ca aplicație nativă pe sisteme desktop, prin Chromium Embedded Framework sau pe sisteme mobile, prin Apache Cordova

Aplicația este construită din mai multe module: autentificare, panou de control, camere de discuție, traducere mesaje, vizualizare termeni regăsiți în ontologie. Prin modulul de autentificare se pot crea noi conturi sau iniția sesiuni persistente cu serverul. După autentificare, se realizează redirectarea către Dashboard (Panoul de control). De aici, un utilizator poate vizualiza camerele de discuție la care are acces, prietenii şi poate adăuga noi camere de discuție.

Datorită creşterii tot mai mari a accesului la Internet prin intermediul dispozitivelor mobile, s-a considerat necesară şi posibilitatea adaptării automate a interfeței la rezoluțiile specifice acestor dispozitive. Astfel, s-a decis că dezvoltarea trebuie să se axeze în primul rând pe dispozitive mobile. În acest mod, sunt accesibile doar funcționalitățile minime necesare prin care un utilizator poate folosi aplicația la capacitate maximă (Figura 3.a). Când aplicația este accesată de pe un ecran mai mare (Figura 3.b), datorită spațiului adițional, sunt prezente mai multe elemente, fiind sporită astfel interacțiunea cu utilizatorul.

Pentru interfață au fost folosite următoarele biblioteci:

- bootstrap (dezvoltat de Twitter). S-a folosit sistemul grid (grilă) pentru a permite adaptarea automată a aplicației pe diferitele rezoluții ale dispozitivelor pe care va fi accesată.
- angular (dezvoltat de Google). S-a folosit pentru a permite dezvoltarea întregii arhitecturi de structurare, navigare şi încărcare dinamică a datelor.



a b

Figura 3. Diferența de prezentare a interfeței utilizator în funcție de rezoluție: a) Rezoluție 320 x480, meniul adițional apare la apăsarea butonului de meniu, b) Rezoluție 1366 x 768, mai multă informație vizibilă

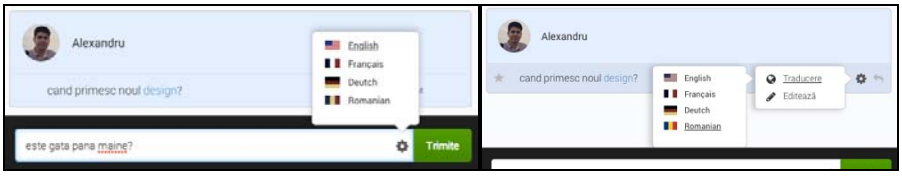
Serverul a fost dezvoltat cu ajutorul următoarelor tehnologii: nodejs, redis, mongodb. Redis este folosit ca mediu distribuit de stocare al sesiunilor utilizatorilor, iar mongodb pentru persistența datelor. Aplicația permite comunicarea între client şi server prin websockets şi trimiterea de notificări în timp real către clienți ca urmare a actualizărilor făcute în modelele de date. Astfel, în momentul în care un utilizator adaugă o nouă resursă (de exemplu, scrie un nou mesaj), acesta este mai întâi salvat în baza de date. Apoi, este transmisă o notificare către toate modelele ce sunt abonate la actualizările resursei tocmai modificate, în cazul de față al mesajului tocmai salvat. Aceste actualizări sunt trimise către celelalte aplicații client ce sunt conectate în acel timp la server. Întrucât aplicația este dinamică, actualizările sunt reflectate automat în regiunile din interfață ce depind de modelele de date modificate.

4.1 Detecția limbii şi traducerea mesajelor

Înțelegerea limbii folosite în discuție este necesară pentru detecția şi corecția automată a greşelilor de ortografie comise de utilizatori. Astfel, participanții la dialog pot transmite un mesaj mai clar şi corect, iar contextul discuțiilor poate fi ulterior mai bine înțeles.

Posibilitatea traducerii mesajelor în alte limbi elimină barierele lingvistice care pot împiedica participarea anumitor persoane la dialog. Oferirea unei modalități cât mai simple de traducere a unui mesaj în limba dorită este o funcționalitate ce va ajuta un număr mai mare de participanți să contribuie activ, dar şi alte persoane să înțeleagă subiectul discutat.

Serverul realizează o detecție a limbii fiecărui mesaj trimis, folosind algoritmul n-gram. Limba detectată este folosită ulterior dacă se doreşte traducerea mesajului sau dacă acesta are greşeli ortografice. În stadiul actual sunt suportate 52 de limbi prin intermediul serviciului web pus la dispoziție de Bing.



a b

Figura 4. Modalitățile de traducere ale unui mesaj oferite de aplicație

Mesajul se poate traduce fie înainte de a fi trimis, fie dacă este prezent în conversație. Mesajul poate fi tradus automat înainte de a fi postat (figura 4.a), sau poate fi selectat orice mesaj din camera pentru traducere ulterioară (figura 4.b).

4.2 Vizualizarea ontologiilor

Într-o cameră de discuție, în momentul în care este introdus un nou mesaj (Figura 5), acesta este analizat şi, dacă acesta se regăseşte în ontologiile predefinite , acesta poate fi vizualizat, folosind biblioteca JavaScript theJit, sub forma unui graf radial generat pe baza relațiilor dintre conceptele ontologiei selectate (Figurile 6 şi 7).

Figura 5. Detecția în ontologie a unui cuvânt din mesaj

Figura 6. Vizualizarea grafului radial în care a fost regăsit cuvântul *Design*

5. Concluzii

Aplicația dezvoltată reprezintă un mediu de comunicare în timp real între utilizatori ce facilitează comunicarea şi colaborarea prin intermediul schimbului de mesaje între 2 persoane (1:1) sau între un utilizator şi un grup (1:N). Termenii din mesaje care sunt regăsiți în ontologii pot fi vizualizați, iar camerele de discuție sunt clasificate pe baza unui algoritm statistic de numărare a frecvenței termenilor în ontologii.

Prin posibilitatea exportării conversațiilor în format XML sau JSON, acestea pot fi prelucrate independent de aplicație. Astfel, aplicația poate constitui un schelet pentru partea de comunicare, iar alți algoritmi pentru analiză şi clasificare pot fi folosiți pe baza arhivei de discuții.

Figura 7. Vizualizarea grafului radial în care a fost regăsit cuvântul *Film*

De asemenea, identificarea limbii folosite într-o conversaţie ajută la o mai bună înţelegere a contextului. Cu cât detecţia este mai bună, se pot oferi sugestii mai precise pentru corecţia ortografică a cuvintelor. Algoritmii statistici prezentaţi, atât cel de corecţie cât şi cel pentru detecţia limbii au o precizie mai bună cu cât volumul iniţial de date de antrenare este mai mare. Pentru corecţie s-au folosit doar date de antrenare în limba engleză, dar principiile sunt aplicabile la nivelul celorlalte limbi suportate.

Combinând aceşti algoritmi şi având un set cât mai mare de ontologii disjuncte la nivelulul cărora se efectuează regăsirea termenilor conversaţiilor purtate, se poate determina subiectul discuţiilor. Existenţa mai multor ontologii nu duce obligatoriu la o clasificare mai bună întrucât elementele comune în mai multe ontologii introduc ambiguitate şi interpretabilitate. Astfel, acestea pot fi clasificate cât mai precis, în funcţie de domeniul din care fac parte ontologiile predominante.

Ca direcţii de dezvoltare ulterioară, se consideră necesară traducerea ontologiilor pentru o mai bună mapare a acestora pe termenii fiecărei limbi, precum şi extinderea setului curent de ontologii folosite. Desigur, utilizarea platformei în scenarii educaţionale integrate în cadrul unei platforme de e-learning este considerată o extensie naturală a eforturilor actuale (Iftene & Rotaru, 2010). Adiţional, avem în vedere două funcţionalităţi suplimentare. Pe de o parte, pentru camerele de discuţii care nu au subiectul clasificat, se va cere clasificarea manuală de către utilizator, facilitând astfel crearea ulterioară a unui algoritm de învăţare automată pentru îmbunătăţirea predicţiei. Pe de altă parte, pornind de la exportul conversaţiei, ne dorim integrarea cu platforma *ReaderBench* (Dascălu et al., 2013; Dascălu, 2014) în vederea evaluării automate a gradului de participare şi a colaborării participanţilor. Astfel, aplicaţia de derulare a conversaţiilor integrată cu platforma *ReaderBench* va permite derularea de multiple experimente centrate pe evaluarea conversaţiilor chat derulate în diverse scenarii educaţionale (spre ex., rezolvarea colaborativă de problemelor, dezbateri).

Mulţumiri

Rezultatele prezentate în acest articol au fost obţinute cu sprijinul Ministerului Fondurilor Europene prin Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007-2013, Contract nr. POSDRU/159/1.5/S/134398.

Referințe

Angular, disponibil online la https://angularjs.org/, accesat la data de 18.06.2014

Apache Cordova, disponibil online la http://cordova.apache.org/, accesat la data de 18.06.2014

Appear.i, disponibil online la https://appear.in/, accesat la data de 18.06.2014

Asana, disponibil online la https://asana.com/, accesat la data de 18.06.2014

Babel Fish, disponibil online la http://www.babelfish.com/, accesat la data de 18.06.2014

Berners-Lee, T., Hendler, J., Lassila, O. (2001), Scientific American*, The Semantic Web*

Bing Translator, disponibil online la http://www.bing.com/translator/, accesat la data de 18.06.2014

Bootstrap, disponibil online la http://getbootstrap.com/, accesat la data de 18.06.2014

Bright, P. (2011), *Microsoft Buys Skype for $8.5 Billion. Why, Exactly?*, Wired, publicat online la http://www.wired.com/2011/05/microsoft-buys-skype-2/, accesat la data de 18.06.2014

Campfire, disponibil online la https://campfirenow.com/, accesat la data de 18.06.2014

Cavnar, W.B., Trenkle, J.M. (1994), *N-Gram-Based Text Categorization*, Environmental Research Institute of Michigan

Chromium Embedded Framework, disponibil online la https://code.google.com/p/chromiumembedded/, accesat la data de 18.06.2014

Cisco, *Cisco's Visual Networking Index Forecast Projects Nearly Half the World's Population Will Be Connected to the Internet by 2017*, The Network Cisco's Technology News Site, disponibil online la http://newsroom.cisco.com/release/1197391/, 2013, accesat la data de 18.06.2014

ConcertChat, disponibil online la http://sourceforge.net/projects/concertchat/), accesat la data de 18.06.2014

Dascălu, M. (2014), *Analyzing discourse and text complexity for learning and collaborating,* Studies in Computational Intelligence. Springer, Switzerland

Dascălu, M., Trăuşan -Matu, S.,Dessus, P. (2013), *Cohesion-based analysis of CSCL conversations: Holistic and individual perspectives,* In 10th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2013), N. Rummel, M. Kapur, M. Nathan and S. Puntambekar Eds. ISLS, Madison, USA, pp. 145–152

Deshpande, Y., Murugesan, S., Ginige, A., Hansen, S., Schwabe, D., Gaedke, M., White, B. (2002), *Web Engineering,* Journal of Web Engineering, 1 (1), pp. 3-17

Google Translate, disponibil online la https://translate.google.com/, accesat la data de 18.06.2014

HipChat, disponibil online la https://www.hipchat.com/, accesat la data de 18.06.2014

Iftene, A., Rotaru, A. 2010. *Îmbunătățirea unui sistem de eLearning folosind elemente specifice unui sistem de tip întrebare-răspuns*. In Revista Română de Interacțiune Om-Calculator (RRIOC), Vol. 3 (2010) Numar special: Interactiune Om-Calculator 2010,

ISSN 1843-4460, Pp. 71-74, Bucuresti, 2-3 Septembrie, 2010.

Internet Working Group, *Internet Relay Chat Protocol*, disponibil online la https://tools.ietf.org/html/rfc1459, accesat la data de 18.06.2014

Jira, disponibil online la https://www.atlassian.com/software/jira, accesat la data de 18.06.2014

Levenshtein, V.I. (1966), "*Binary codes capable of correcting deletions, insertions, and reversals*".Soviet Physics Doklady, 10 (8), pp. 707–710

McAfee, A., Brynjolfsson, E. (2012), *Big Data: The Management Revolutio*, Harvard Business Review, publicat online la http://hbr.org/2012/10/big-data-the-management-revolutionar, accesat la data de 18.06.2014

Mongodb, disponibil online la www.mongodb.org/, accesat la data de 18.06.2014

Mozilla Developer Network, *WebRTC*, disponibil online la https://developer.mozilla.org/en-US/docs/Web/Guide/API/WebRTC), accesat la data de 18.06.2014

Nodejs, disponibil online la http://nodejs.org/, accesat la data de 18.06.2014

Norvig, *How to Write a Spelling Corrector*, disponibil online la http://norvig.com/spell-correct.html, accesat la data de 18.06.2014

Powell, T., Jones, D., Cutts, D. (1998), *Web Site Engineering: Beyond Web Page Design*, Prentice Hall

Quora, disponibil online la http://www.quora.com/, accesat la data de 18.06.2014

Reddis, disponibil online la http://redis.io/, accesat la data de 18.06.2014

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. In 14th Int. Joint Conf. on Artificial Intelligence (IJCAI'95) Morgan Kaufmann, Montreal, Canada, 448–453.

Single Page Application, disponibil online la http://www.johnpapa.net/pageinspa/, accesat la data de 18.06.2014

Slack, disponibil online la https://slack.com/, accesat la data de 18.06.2014

Stackoverflow, disponibil online la http://stackoverflow.com/, accesat la data de 18.06.2014

Stahl, G., Group cognition. Computer support for building collaborative knowledge. MIT Press, Cambridge, MA, 2006.

Stone, B. (2014), *Facebook Buys WhatsApp for $19 Billion*, Bloomberg Businessweek, publicat online la http://www.businessweek.com/articles/2014-02-19/facebook-acquires-whatsapp-for-19-billion, accesat la data de 18.06.2014

theJit, disponibil online la http://philogb.github.io/jit/), accesat la data de 18.06.2014

Trăuşan -Matu, S., Rebedea, T. (2010), *A polyphonic model and system for inter-animation analysis in chat conversations with multiple participants*, In 11th Int. Conf. Computational Linguistics and Intelligent Text Processing (CICLing 2010), A.F. Gelbukh Ed. Springer, New York, pp. 354–363

Trello, disponibil online la https://trello.com/, accesat la data de 18.06.2014