Revista Română de Interacțiune Om-Calculator 8 (3) 2015, 173-194 © MatrixRom

Analiza comentariilor publicate pe platforma YouTube folosind tehnici de prelucrare a limbajului natural

Iulian Radu, Traian Rebedea

Universitatea Politehnica din Bucureşti, Facultatea de Automatică şi Calculatoare
Splaiul Independenței nr. 313, Sector 6, Bucureşti 060042
E-mail: iuliradu@gmail.com, traian.rebedea@cs.pub.ro

Rezumat. Publicarea de comentarii este o formă de interacțiune care a devenit parte din fenomenul de internet social. Comentariile întăresc conceptul de interacțiune online, însă vizualizarea acestora devine dificilă sau chiar imposibilă pentru videoclipurile şi alte articole online care au milioane de comentarii înregistrate. Lucrarea prezintă o metodă prin care se pot identifica cele mai relevante comentarii publicate pentru un videoclip distribuit pe platforma YouTube, oferind în acelaşi timp o analiză detaliată a atributelor care sunt relevante pentru cele mai importante comentarii. Analiza efectuată arată că videoclipurile din categorii diferite au atribute relevante diferite, constatându-se diferențe destul de mari între unele categorii (de ex. Muzică vs. Educație).

Cuvinte cheie: regăsirea informației, prelucrarea limbajului natural, internet social, analiză comentarii, algoritm de relevanță, ordonare rezultate

Introducere

Pentru a defini într-un cuvânt fenomenul de web social se poate folosi termenul de *interacțiune.* Dacă acum zece ani utilizatorii de internet vizitau în special paginile web pentru a accesa sau pentru a descărca informații, acum comunicarea între utilizatori şi serviciile oferite este mult mai interactivă, astfel încât paginile statice se întâlnesc din ce în ce mai rar.

Comentariile publicate pe platforma YouTube (http://www.youtube.com) sunt un bun exemplu de interacțiune şi de web social în continuă schimbare. Vizitatorii accesează videoclipurile, apasă pe butonul Like sau Dislike, partajează videoclipul, adaugă comentarii sau răspund la comentariile existente. Această categorie de funcționalități este întâlnită din ce în ce mai des, iar rolul acesteia este de a oferi utilizatorilor posibilitatea să-şi exprime părerea referitor la un anumit subiect, iar în acest fel conținutul este

imbogăţit în mod dinamic, comentariile crescând sau micşorând valoarea unei pagini în funcţie de o serie complexă de factori.

În cazul YouTube, există videoclipuri care au înregistrat peste două miliarde de vizualizări , ceea ce înseamnă că pot exista aproximativ cinci milioane (Chatzopoulou, et al., 2010) de comentarii adăugate de către utilizatori pentru un singur videoclip. În aceste situaţii, în care numărul de comentarii adăugate este mult prea mare pentru a putea fi parcuse de către utilizatori, este necesară implementarea unor metode prin care să se poată extrage cele mai relevante comentarii. Lucrarea îşi propune să trateze acest aspect, pornind de la indexarea comentariilor, procesarea şi modelarea acestora până la afişarea rezultatelor concrete prin aplicarea unor filtre.

Secţiunile 2 şi 3 fac o scurtă introducere în contextul platformelor sociale, modului de calcul al popularităţii videoclipurilor pe platforma YouTube şi domeniul de interacţiune în internetul social prin intermediul publicării şi vizualizării de comentarii.

Secţiunea 4 detaliază implementarea tehnică a soluţiei pentru indexare, modelare şi analiză a comentariilor, fiind detaliate componetele logice ale sistemului, dar şi detalii specifice funcţionalităţilor cheie. Secţiunea descrie atât modul de lucru pentru colectarea şi procesarea comentariilor, cât şi modul de lucru pentru activarea sau dezactivarea filtrelor şi detalii despre filtrele disponibile în aplicaţie.

Secţiunea 5 conţine evaluarea rezultatelor atât la nivel general, pe baza celor 750.000 de comentarii indexate în aplicaţie, cât şi detalii specifice pentru analiza efectuată la nivel de comentariu prin aplicarea filtrelor de sortare pentru videoclipuri din diferite categorii. Analiza efectuată la nivel de comentariu este realizată prin compararea rezultatelor obţinute prin aplicarea filtrelor cu rezultatele obţinute prin adnotarea manuale de către utilizatori.

Algoritmii de filtrare vizează inclusiv factorii emoţionali cu privire la comentariile preluate, pe baza unor servicii specifice –Textalytics si Open Calais. În acest mod, filtrele aplicate pot oferi un grad mai mare de relevanţă, prin aplicarea corelată a acestor factori de filtrare.

YouTube pune la dispoziţie un număr limitat de comentarii prin intermediul YouTube API ,iar în cadrul acestei lucrări a fost dezvoltat un

sistem ce permite urmărirea videoclipurilor virale[1], astfel încât să se poată construi în timp o bază de date locală cu toate comentariile adăugate pentru un anumit videoclip. Deținând aceste date, se poate efectua inclusiv o analiză temporală a comentariilor relevante.

Cercetări similare în domeniu

Platforma YouTube a fost lansată în anul 2005 iar în primul an de la publicare au fost adăugate peste 65.000 de videoclipuri. Acesta a înregistrat o creştere foarte mare de la an la an, astăzi fiind cea mai mare platformă de partajare a videoclipurilor din lume, înregistrând un procent semnificativ (Anon., 2013) din traficul total de date din internet. Din acest motiv, analiza modului în care utilizatorii interacționează cu această platformă poate oferi informații relevante pentru tendințele globale în web-ul social.

Ținând cont de vechimea acestei platforme, înainte de anul 2008 nu au fost realizate studii notabile referitoare la popularitatea videoclipurilor, comentarii sau interacțiunea utilizatorilor cu plaforma, primul şi cel mai cuprinzător având ca obiectiv analiza comentariilor fiind făcut în anul 2010 (Sierdorfer, et al., 2010) folosind concepte similare de analiză cu cele tratate în această lucrare, cum ar fi lista de termeni pozitivi, termeni negativi sau integrarea cu SentiWordNet (Esuli & Sebastiani, 2006) pentru a afla informații referitoare la obiectivitatea sau subiectivitatea comentariilor. În ansamblu, au fost abordate mai multe tipuri de criterii, similar cu lucrarea de față, iar analiza efectuată oferă un punct de plecare solid pentru analiza relevanței comentariilor ținând cont de un număr cât mai mare de factori identificați.

Analiza relevanței comentariilor constă într-o separare a factorilor pe mai multe nivele, în funcție de impactul acestora. Pot fi factori macro, ce pot fi analizați pe baza unui set foarte mare de date, sau factori specifici, orientați către structura şi informațiile disponibile la nivel de comentariu, aşa cum se specifică în cadrul raportului (Hsu, et al., 2009) pentru criteriile de complexitate şi gradul de informare.

---

[1] Videoclipuri care sunt partajate de la utilizator la altul cu o rată suficient de mare astfel încât numărul vizualizărilor acestora cre□te constant

În cadrul celui mai amplu proiect pentru analiza informațiilor asociate videoclipurilor publicate pe YouTube (Chatzopoulou, et al., 2010), au fost indexate 40 de milioane de videoclipuri, 41 de milioane de comentarii, însă analiza datelor este orientată către popularitatea videoclipurilor şi nu în mod special pe analiza comentariilor. Din acest motiv, raportul dintre numărul de comentarii indexate şi numărul de videoclipuri este de aproxativ un comentariu pentru un videoclip îndexat, diferit față de implementarea aferentă acestei lucrări, unde media este de 744 de comentarii pentru un videoclip.

Lucrarea de față se diferențiază de celelalte lucrări similare prin faptul că elementul cheie în cadrul acestei analize este videoclipul şi se urmăreşte stabilirea unui set de reguli de ordonare ținând cont de comentariile specifice ale unui videoclip ci nu doar pe baza analizării unui corpus de comentarii.

Într-o lucrare anterioară (Chatzopoulou, et al., 2010) se analizează videoclipurile publicate pe YouTube tot din perspectiva popularității acestora, pentru identificarea factorilor de ordin macro care stau la baza popularității în site-uri care au conținut generat de utilizatori.

Există cateva publicații ce tratează acest subiect, chiar şi din perspectiva socială analizând modul în care utilizatorii reacționează din punct de vedere emoțional atunci când citesc diverse comentarii sau când alți utilizatori postează comentarii la filmările proprii (Lange, 2007). A fost realizată o clasificare a tiparelor de utilizator şi o analiză a părerilor acestora despre modul în care funcționează comentariile. Unii utilizatori sunt deranjați de către cei care adaugă comentarii jignitoare, însă alți nu au nici o problemă cu acest lucru (Lange, 2007). Unii dintre cei mai populari utilizatori ai YouTube au propus o soluție prin care comentariile să poată fi adăugate în funcție de scorul personal, care se calculează pe baza notelor obținute pentru comentariile tale, de la ceilalți utilizatori. Cu toate astea, o astfel de soluție nu poate fi generalizată pentru că este strict dependentă de contextual socio-cultural în care se aplică.

Un alt domeniul în care s-a ridicat problema relevanței comentariilor este acela al motoarelor de căutare (Junqueira & Plachouras, 2007). Pentru că în general cuvintele cheie din căutări conțin termeni ce se bazează pe context, nu există o compatibilitate cu termenii din comentarii ce poate fi exploatată. Cu alte cuvinte, comentariile conțin puține informații ce ar putea fi folositoare motoarelor de căutare. Cu toate astea mai sunt şi alte probleme în ceea ce priveşte căutarea, deoarece comentariile se generează în timp şi

este nevoie de o perioadă considerabilă până ca un obiect să poată avea asociat un număr suficient de comentarii astfel încât aceastea să fie relevante pentru căutare.

În (Chatzopoulou, et al., 2010) se efectuează o corelație între valorile a diferiți parametri la nivel global, iar datele obținute arată faptul că variabilele calculate sunt corelate şi pot fi folosite pentru calcularea de predicții, folosind formule de regresie liniară multiplă. De asemenea, analizând un număr de peste 40 de milioane de videoclipuri, s-a obținut o corelare între numărul de vizite asociate unui comentariu şi numărul de interacțiuni efectuate de utilizatori. Pe baza acestor date s-a obținut un indicator de interacțiune la 400 de vizite, interacțiunea însemnând adăugarea unui comentariu, utilizarea *like/dislike*, adăugare în lista de comentarii favorite sau urmărirea canalului. Prin analiza unui număr arbitrar de videoclipuri publicate pe YouTube, de preferat videoclipuri cu o vechime şi expunere relative mari, se poate confirma faptul că numărul de interacțiuni este de aproximativ 400 de ori mai mic față de numărul de vizite.

Implementarea soluției

Descrierea modului de funcționare

Întreg procesul de analiză a comentariilor se desfăşoară cu ajutorul unei aplicații web, având funcționalități pentru întreg fluxul de lucru, începând cu identificarea videoclipurilor virale, relevante pentru analiză, colectarea şi stocarea datelor, inclusiv analiză a relevanței şi afişare a datelor statistice. Prima componentă constă în identificarea automată a videoclipurilor populare şi urmărirea acestora. Aplicația detectează automat videoclipurile populare la nivel de țară, zilnic, şi le adaugă în lista de videoclipuri monitorizate periodic. Componenta de urmărire a videoclipurilor verifică la un interval de timp stabilit, apariția de noi comentarii pentru fiecare videoclip, iar în cazul în care se găsesc comentarii noi acestea sunt salvate în baza de date.

Graficul de mai jos oferă informații privind execuția metodelor de colectare a datelor, la nivel de oră. Acesta este folosit pentru identificarea situațiilor netratate, în procesul de colectare.

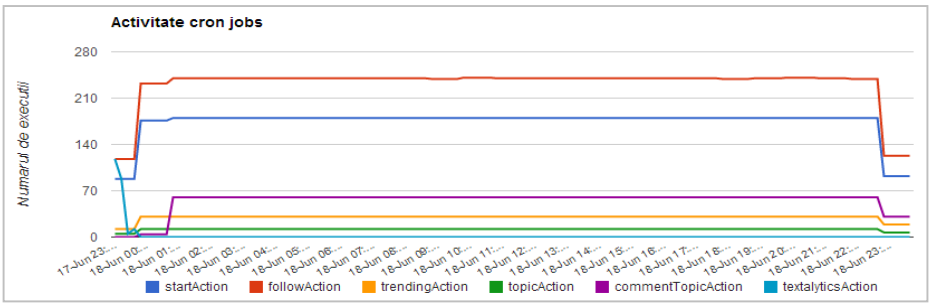

Figura 1 – Grafic pentru monitorizarea activităților curente

Axa Y reprezintă intervalul orar, la nivel de oră, iar coloana X reprezintă numărul de execuții ale metodei într-un interval de o oră. Graficul afişează execuţiile din ultimele 24 de ore.

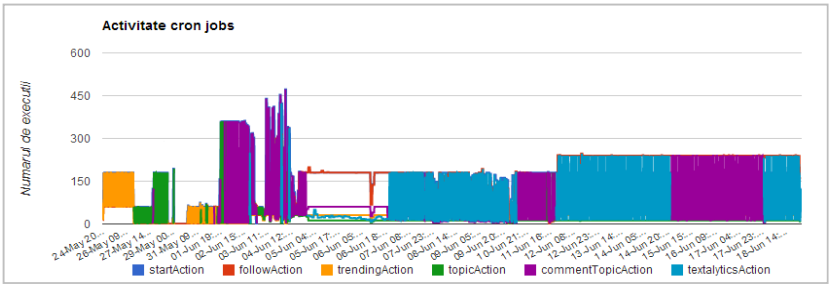

Figura 2 – Activitate Cron Jobs pentru întreaga perioadă

Graficul din *Figura 2* este calculat pe baza unui număr total de 193915 execuţii a metodelor de indexare a comentariilor. În grafic sunt afişate date atât din perioada de dezvoltare şi testare a aplicaţiei cât şi din perioada de colectare a comentariilor. Se poate observa faptul că în prima parte numărul de execuţii oscilează, însă în a doua parte acesta a fost stabilizat.

Urmărirea automată a videoclipurilor populare

Pentru a identifica videoclipurile populare, aplicaţia urmăreşte automat o serie de newsfeed-uri în limba engleză şi indexează în baza de date toate videoclipurile găsite. Se urmăresc videoclipuri pentru următoarele ţări: Statele Unite ale Americii, Marea Britanie, Canada şi lista generală cu videoclipuri populare. Pentru a menţine un număr relativ mic de

videoclipuri urmărite, aplicația elimină videoclipurile care au o frecvență foarte mică de adăugare a comentariilor.

Folosind aceste criterii de indexare şi de ştergere a videoclipurilor din coada de urmărire, a fost păstrat un numărul relativ constant de aproximativ 100 de videoclipuri urmărite la un anumit moment de timp.

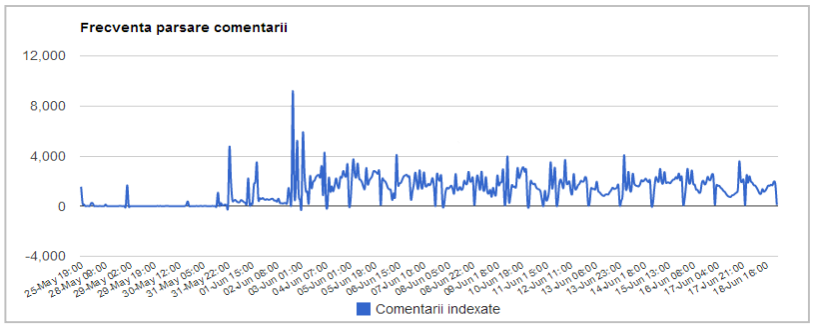

Figura 3 – Frecvența de indexare a comentariilor la nivel de oră

În *Figura 3*, Axa X afişează numărul total de comentarii indexate la nivel de oră, pentru toate videoclipurile urmărite. Graficul este calculat pe o perioadă de aproximativ 25 de zile, în care au fost indexate aproximativ 630.000 de comentarii.

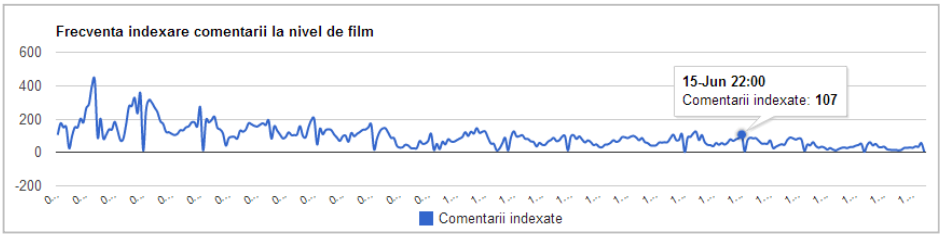

Figura 4 – Frecvența de indexare comentarii la nivel de film

Graficul din *Figura 4* este folosit pentru monitorizarea frecvenței de indexare a comentariilor la nivel de oră. Pentru acest videoclip au fost indexate în total 28.000 de comentarii.

Analiza comentariilor folosind servicii externe

Analiza a fost efectuată atât la nivel local, pe baza textului obținut prin preluarea comentariilor, cât şi pe baza unor servicii externe, cum ar fi Calais

sau Textalytics. Componenta de analiză dezvoltată în cadrul acestei lucrări trimite textul fiecărui comentariu către cele două servicii, iar aceastea furnizează informații referitoare la obiectivitatea, subiectivitatea, nivelul de pozitivism, negatisivm, nivelul de ironie, al comentariilor. Datele obținute sunt stocate local şi sunt folosite în algoritmii de calculare a relevanței. Analiza comentariilor cu ajutorul acestor servicii se face asincron, pe baza unei cozi de aşteptare, pentru toate comentariile unui videoclip.

Tipuri de informații colectate

Aplicația indexează videoclipurile şi comentariile asociate acestora folosind Cron Jobs(planificator de executare a unor acțiuni, la un anumit interval de timp) la nivelul serverului de aplicație. Pentru indexarea informațiilor, au fost definite următoarele actiuni, fiecare colectând un anumit tip de informații:

- **commentstopicAction** – indexează subiectele la nivel de comentariu folosind serviciul Textalytics;
- **followAction** – actualizează comentariile pentru videoclipurile urmărite;
- **profileAction** – indexează informații referitoare la profilul utilizatorilor care au adăugat comentarii;
- **similarityAction** – calculează similaritatea între comentarii şi metadatele videoclipului, folosind algoritmul Cosine Similarity (Manning, et al., 2008);
- **startAction** – se execută încontinuu pentru indexarea tutoror comentariilor unui videoclip;
- **trendingAction** – se execută la un interval de 5 minute şi verifică listele cu videoclipuri populare;
- **userFeed** – indexează informații referitoare la UserFeed-ul fiecărui utilizator care a adăugat comentarii**Invalid source specified.** şi efectuează o analiză referitoare la reputația utilizatorului (Han, et al., 2009);
- **TextalyticsAction** – indexează informații privind topicele din comentarii, etichetele şi informații privind clasificarea emoțională;

Analiza comentariilor

Sistemul de ordonare

Aşa cum motoarele de căutare folosesc algoritmi pentru a clasifica paginile în funcție de importanță, se va folosi un concept similar pentru a ordona comentariile. În cazul comentariilor, relevanța se va calcula pe baza unor

criterii clar definite, care vor avea o pondere proporțională cu aportul de relevanță pe care îl aduce criteriul respectiv.

În (Sierdorfer, et al., 2010) se abordează un sistem de modelare global, la nivel de domeniu, unde pe baza unui set mare de date se calculează o listă cu comentariile "cele mai pozitive" şi „cele mai negative", iar comentariile nou adăugate sunt încadrate în una dintre aceste două categorii. Desigur, este un criteriu de ordonare care reflectă reacția comunității în fața diverselor cuvinte, însă nu este o măsură suficientă pentru sortare.

O abordare mai complexă şi în acelaşi timp mai promițătoare este prezentată în (Hsu, et al., 2009), unde modelarea se face în funcție de un număr mult mai mare de criterii care sunt atât orientate către utilizator, cum ar fi analiza autorității utilizatorului în comunitate, analiza activității utilizatorului într-o anumită categorie, cât şi criterii orientate către conținut, cum ar fi lungimea, complexitatea, informativitatea, subiectivitatea şi unicitatea comentariilor. Aceste criterii sunt concentrate pe anumite subiecte, iar combinația lor creează un criteriu general mai puternic şi mai apropiat de modelul de filtrare al comunității utilizatorilor.

Analizând cele două clase de criterii se observă faptul că ambele dau rezultate, iar dacă ar fi folosite simultan s-ar putea realiza o filtrare cu precizie ridicată ce poate permite modelarea datelor într-un mod complex şi apropiat de modul de gândire şi filtrare al utilizatorilor.

Reguli de ordonare

În cadrul sistemului de față ordonarea comentariilor relevante se va face în funcție de două clase de criterii. O clasă globală care analizează un set de date arbitrar, general şi furnizează informații la nivel de comunitate despre preferințele utilizatorilor în ceea ce priveşte comentariile şi o clasă locală, ce ține cont de informațiile individuale ale unui videoclip. Pentru clasa locală, se colectează informații privind detaliile filmărilor, cum ar fi titlul, descrierea, tag-urile, sau numărul de vizualizări.

În urma analizări gradului de acceptanță al comentariilor (Sierdorfer, et al., 2010) s-au realizat două tabele, cu "cele mai acceptate", respectiv "cele mai respinse" cuvinte cheie, în funcție de voturile pe care aceastea le-au primit de la comunitate. Criteriul principal în clasificarea globală se realizează prin filtrarea comentariilor în funcție de reacția comunității la

diferite cuvinte, ținând cont de gradul de acceptare sau de respingere al acestora (Sierdorfer, et al., 2010).

Ordonarea comentariilor în funcție de relevanță se calculează în două etape. În faza inițială se aplică clasa de filtre globale care generează o ordonare rudimentară, brută, după care intervin criteriile locale care rafinează rezultatele. Filtrele globale funcționează similar cu algoritmii pentru identificarea comentariilor de tip spam, pentru blog-uri, însă la o scară simplificată (Anon., 2005). Compoziția ranking-ului general se face modularizat pentru criteriile locale, iar în urma aplicării algoritmului pentru ranking-ul global acesta este alterat secvențial atunci cand se aplică criteriile locale. Fiecare criteriu local este caracterizat printr-un "Coeficient de importanță", o pondere reprezentată de o valoare numeric subunitară. În implementarea curentă a modului de agregare a clasificării furnizate de fiecare comentariu, ponderea fiecărui filtru este ajustată manual, însă implementarea unui calcul automat al ponderii fiecărui filtru ar putea îmbunătăți semnificativ precizia rezultatelor.

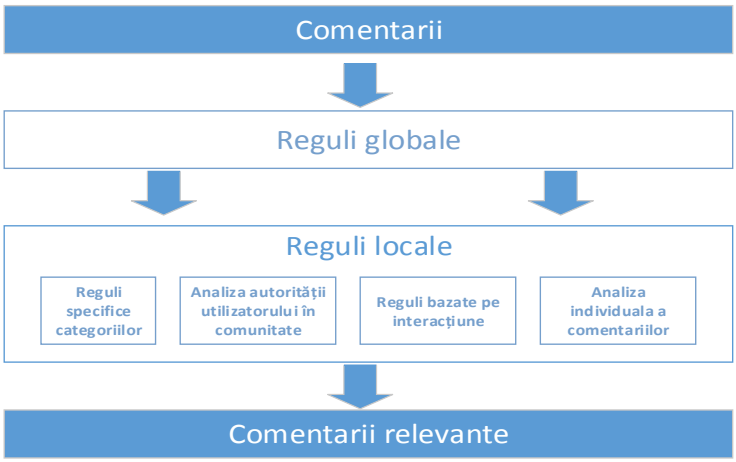

Figura 5. Ordinea şi nivelul de aplicare al claselor de criterii

Criteriile locale sunt modularizate, sunt adăugate sub forma unor componente (plugin-uri) şi li se poate măsura performanța în mod independent. Pe lângă "Coeficientul de Importanță" al fiecărui criteriu acesta calculează pentru fiecare comentariu asociat unei filmări propriul grad de relevanță. Aşasar la nivel de comentariu, atunci când criteriul este aplicat acesta influențează ranking-ul curent, fiind o funcție de parametrii

(*Coeficient de Importanță*, *CommentBoost*), unde parametrul *CommentBoost* reprezintă ranking-ul local al criteriului, care se aplică în ranking-ul general.

Pentru ordonarea comentariilor se foloseşte un indicator de relevanță, calculat de fiecare filtru şi normalizat în intervalul [0,1], iar pentru calcularea rezultatelor agregate pentru mai multe filtre activate simultan se calculează o medie ponderată. Rezultatele obținute pe baza aplicării criteriilor de ordonare sunt analizate la nivel de categorie de videoclupuri, întrucât o analiză globală nu este la fel de relevantă deoarece intervin aspecte sociale şi culturale în funcție de categorie şi zona demografică.

În ceea ce priveşte adaptabilitatea sistemului, au fost analizate modalități pentru a ajusta coeficientul de importanță a fiecărui filtru individual în funcție de performanța obțiuntă prin indicatorul de Average Precision (Manning, et al., 2008). Inițial vor fi agreate doar criterii locale, urmând ca în funcție de performanța acestora să fie adăugate altele noi. Aceastea sunt clasificate în funcție de tipul datelor pe baza cărora se face analiza. Pot fi criterii bazate pe activitatea utilizatorilor (Hsu, et al., 2009), criterii bazate pe conținut (Sierdorfer, et al., 2010) şi criterii generale care pot combina mai multe tipuri de date, inclusiv statistici referitoare la interacțiunea generală şi informații colectate pe baza criteriilor globale.

Criterii dependente de autoritatea utilizatorului

Se presupune că relevanța comentariilor este dependentă de utilizator, de autoritatea pe care acesta o are în cadrul comunității şi de modul în care comunitatea primeşte sau respinge comentariile acestuia. Aşadar se poate crea un model al utilizatorului pe baza căruia putem aplica criterii locale:

- **Numărul de comentarii postate** – se analizează numărul de comentarii postate de către utilizator la nivel global cât şi la nivel de categorie;
- **Vechimea utilizatorului** – acest criteriu ține cont de vechimea contului de utilizator;
- **Activitatea pe categorie** – se calculează procentul de comentarii publicate la filmări din aceiaşi categorie în raport cu numărul total de comentarii publicate de către utilizator. Cu cât acesta a publicat mai multe comentarii într-un anumit domeniu se consideră că expertiza sa în domeniul respectiv este mai relevantă;
- **Nivelul de acceptanță în comunitate** – pentru fiecare utilizator se analizează modul în care comentariile sale sunt votate de către ceilalți

utilizatori. Dacă media de acceptare se apropie de una dintre extreme (acceptat sau respins), probabilitatea ca şi comentariile care nu au primit voturi să se îndrepte către una dintre cele două extreme este ridicată;

Criterii dependente de conținut

Această categorie de criterii locale analizează în mod exclusiv conținutul comentariilor, calitatea acestora şi încearcă să imite modul în care utilizatori clasează comentariile din punct de vedere cognitiv. Este cea mai importantă categorie de criterii, întrucât tratează în mod direct aspectele emoționale, sociale care îi fac pe utilizatori să decidă dacă un comentariu este relevant sau nu.

- **Lungimea comentariului** - criteriul măsoară numărul de cuvinte conținute în comentariu;
- **Complexitatea comentariului** - se calculează pe baza entropiei cuvintelor din comentariu. Pe baza formulei de mai jos (Hsu, et al., 2009) unde pentru componenta c_j cu un număr de fiecare cuvânt are frecvența p_i.

$$entropy(c_j) = \frac{1}{\lambda}\sum_{i=1}^{n} p_i[log_{10}(\lambda) - log_{10}(p_i)]$$

;

- **Unicitatea conținutului** - se calculează unicitatea textului unui comentariu în comparație cu celelalte comentarii ale unui clip video. Unicitatea unui comentariu c_i se calculează folosind o variație standard TFIDF (Sierdorfer, et al., 2010) (term frequency – inverse document frequency) pentru colectarea datelor, iar gradul de unicitate al unui comentariu este dat de suma tuturor coeficienților pentru fiecare cuvânt unic în cadrul unui comentariu, astfel:

$$inform(c_j) = \sum_{t_i \in c_j} \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

;

- **Gradul de apartenență la categorie** - acest criteriu calculează similitudinea între alte comentarii postate de acelaşi utilizator în alte categorii;
- **Badwords** - identificarea comentariilor ce conțin cuvinte neadecvate folosind liste identificate în diverse surse online.

Sistem comparativ

Sistemul comparativ are rolul de a evalua performanța fiecărui criteriu de filtrare în parte, astfel încât, acesta permite vizualizarea în paralel a rezultatelor pentru criteriile de filtrare aplicate unei liste de comentarii, plus

criteriul ce necesită analiza. De exemplu, dacă avem trei criterii de filtrare deja testate şi dorim adăugarea unuia nou, vom compara rezultatele dupa ce au fost aplicate criteriile 1, 2 şi 3, cu rezultatele pentru care s-au aplicat toate cele 4 criterii. In acest mod, exista posibilitatea de a compara clar rezultatele obținute de criteriul numarul 4, iar daca este nevoie se pot ajusta valori constante care au fost folosite în aplicație, care, în cel mai frecvent caz, depind de natura socială şi categoria comentariilor. (Anon., 2007)

Adnotare manuală a comentariilor

Are rolul de a compara rezultatele obținute prin aplicarea filtrelor cu adnotarea efectuată manual, de către utilizator. Un număr arbitrar de utilizatori votează comentariile unui videoclip, folosind o scară de valori de la 1 la 4, acordându-se valoarea 1 pentru comentariile irelevante, până la valoarea 4 pentru comentariile cele mai relevante şi informative.

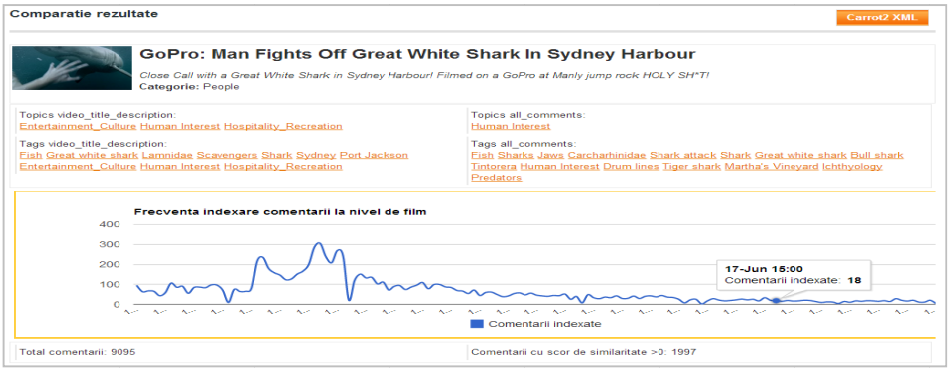

Figura 6 – Videoclip urmărit timp de 6 zile, 9000 de comentarii indexate

Măsurarea performanței folosind Average Precision

Pentru compararea rezultatelor obținute prin aplicarea filtrelor cu adnotarea manuală a comentariilor, s-a folosit formula pentru Mean Average Precision (Manning, et al., 2008). La nivel de comentariu se calculează valoarea pentru precizie, însă este nevoie de o valoare la nivel de listă de documente pentru analiza calității criteriului de analiză.

$$\text{AveP} = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{\text{number of relevant documents}}$$

unde rel(k) este o funcție care returnează valoarea 1 în cazul în care comentariul de pe poziția k este relevant, conform adnotării manuale şi 0 altfel, P(k) este precizia calculată pentru primele k comentarii obținute.

Arhitectura soluției

General

Din punct de vedere logic, sistemul este alcătuit din componenta de colectare a datelor, responsabilă pentru indexarea videoclipurilor şi a datelor asociate, componenta de analiză a datelor, cea care procesează textul în atomi lexicali şi îl analizează prin intermediul serviciilor externe *OpenCalais* şi *Textalytics*.

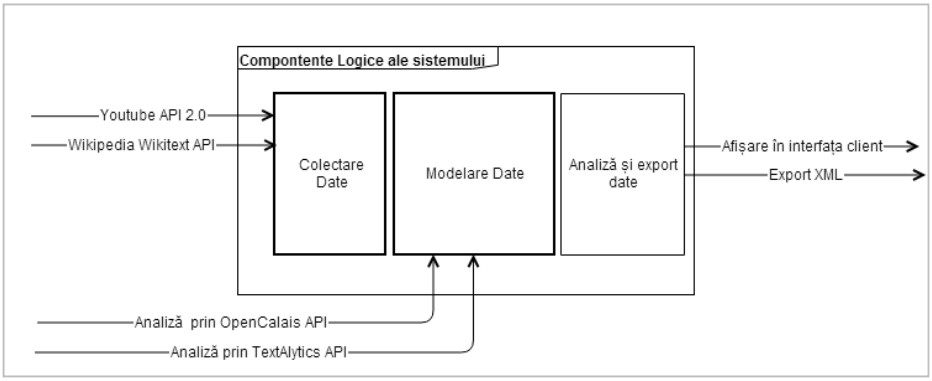

Figura 7 – Arhitectura logică a soluției

Interfața de modelare date

Este folosită ca suport pentru colectarea informațiilor referitoare la comentarii de pe YouTube folosind YouTube Data API. Interfața de modelare a datelor integrează două servicii pentru analiza textului comentariilor, Textalytics şi Open Calais. Acestea servicii furnizează metadate la nivel de comentariu, Open Calais identifică categorii şi subiecte iar Textalytics identifică suplimentar subiectivismul, ironia şi negativismul. Toate comentariile luate în considerare în faza de analiză au fost procesate pe baza acestor servicii, iar informațiile obținute au fost incluse în criteriile de ordonare dezvoltate.

Interfața de analiză

Aceasta permite o analiză detaliată a modului în care funcționează criteriile locale de filtrare deoarece afişează pentru fiecare comentariu în parte modul în care a fost compus scorul de ranking total. În plus, există posibilitatea analizării rezultatelor fiecărui criteriu în parte, afişând două liste de comentarii în paralel, pe două coloane, una când filtrul este activat şi cealaltă când acesta este inactiv.

Evaluarea soluției

Analiza generală a datelor colectate

Modulul de colectare indexează date cu o frevență cuprinsă între 200 şi 2500 comentarii pe oră în funcție de comentariile publicate la nivel de videoclip. În *Tabelul 1* sunt furnizate informații cantitative referitoare la numărul de comentarii indexate şi analiza efectuată pe baza acestora.

Tabel 1 - Informații statistice referitoare la modulul de colectare a datelor

Numărul total de comentarii indexate (în termen de 60 de zile)	761000
Numărul total de videoclipuri indexate	1021
Numărul mediu de videoclipuri urmărite concomitent	100
Numărul total de adnotări manuale adăgate de utilizatori	2308
Numărul maxim de comentarii indexate pentru un videoclip	31433
Numărul total de comentarii analizate cu Textalytics	36339
Numărul total de etichete obținute prin OpenCalais	19440
Numărul total de articole Wikipedia indexate	4842
Numărul de execuții Cron	20000

Open Calais – analiză taxonomie

Taxonomia din Figura 8 este calculată folosind un număr total de 16796 etichete obținute din *Open Calais*, pe baza analizei a aproximativ 623.000 de comentarii. Această furnizează informații privind categoriile de conținut publicate pe Youtube, dintr-o perspectivă externă, doar din analiza textului conținut în comentariile publicate pentru videoclipuri.

Figura 8 – Taxonomie de etichete Calais, pentru toate comentariile indexate

Analiza emoțională

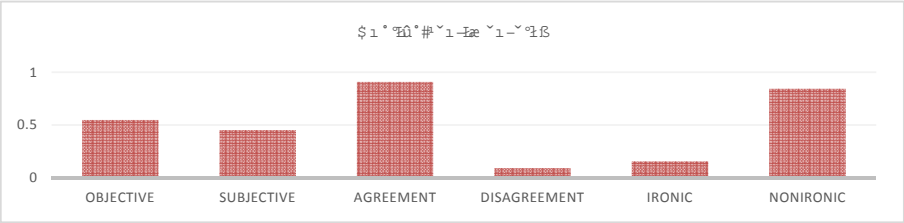

Figura 9 – Analiza sentimentelor folosind Textalytics

Graficul din *Figura 9* a fost calculat pe baza unui număr total de 19.000 de comentarii indexate pentru videoclipuri diferite. Se poate observa faptul că domină comentariile non-ironice și cu un nivel ridicat de aprobare. Aceste rezultate au fost calculate folosind Textalytics, însă pentru obținerea unei imagini de ansamblu mai precise este necesar să se efectueze o astfel de analiză pentru întreaga bază de date, de aproximativ 750.000 de comentarii.

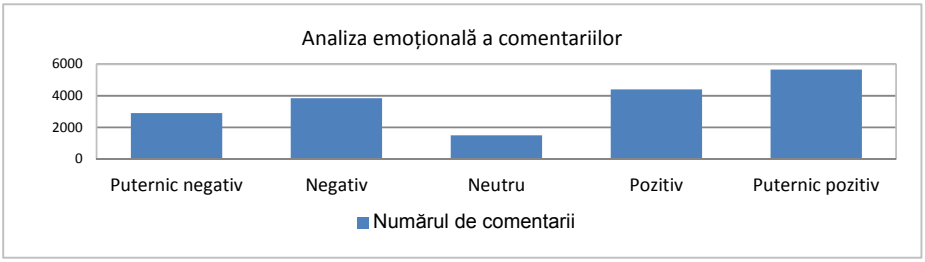

Figura 10 – Statistică privind analiza emoțională a comentariilor indexate

Sistemul analizează fiecare comentariu indexat din punct de vedere emoțional prin intermediul unui serviciu web, acesta furnizând un scor

încadrat în intervalul [-1,1] corespunzător etichetelor din graficul *Analiza emoțională a comentariilor*. Informațiile furnizate de acest serviciu sunt transformate într-un indicator pentru calcularea comentariilor relevante.

Evaluarea filtrelor pentru identificarea comentariilor relevante

În etapa de adnotare manuală a comentariilor s-a observat faptul că există o corelație între categoria din care face parte videoclipul şi lungimea comentariilor adăugate pe baza analizei numărului de tokeni indexați la nivel de comentariu şi din acest motiv s-a efectuat o clasificare a videoclipurilor în funcție de categorie.

Analiza comentariilor indexate în baza de date a arătat faptul că există o diferență semnificativă între lungimea medie a comentariilor şi categorie, existând diferențe de până la trei ori mai mult între anumite categorii.

Tabel 2 – Numărul mediu de atomi lexicali la nivel de comentariu, după categoria videoclipurilor

Categorie	Numărul mediu de atomi lexicali la nivel de comentariu	Numărul de comentarii din categorie
Filme si divertisment	11.83	231
Nonprofit	9.01	252
Călătorii	7.57	869
Ştiri şi politică	6.53	8462
Social	6.50	65540
Educație	6.15	9673
Film	4.94	27894
Tehnologie	4.63	19258
Ghiduri	4.57	11575
Divertisment	4.52	195429
Comedie	4.38	63807
Automobile	4.35	994
Sport	4.29	23195
Jocuri	3.98	191169
Muzică	3.92	126325
Animale	3.55	1031

Întrucât lungimea comentariilor şi gradul de interes față de comentarii al utilizatorilor este corelat cu categoria videoclipurilor, este necesară o analiză pentru videoclipuri din fiecare cele trei categorii definite mai sus. Cum filtrele pentru calcularea relevanței comentariilor depind de numărul de cuvinte, cu cât un comentariu are mai multe cuvinte, cu atât se pot obține mai multe informații despre acesta. Aşadar, filtrele pot avea un impact mai

mare sau mai mic asupra setului de comentarii asociat unui videoclip în funcţie de categoria acestuia. În cadrul unui alt studiu, unde au fost colectate 41.1 milioane de comentarii (Chatzopoulou, et al., 2010), s-a efectuat o analiză a distribuţiei comentariilor pe categorii, iar videoclipurile din categoriile Music şi Entertainment reprezintă 45% din totalul videoclipurilor salvate.

Categorie: Entertainment

Analiza a fost efectuată pentru un număr de aproximativ 200 de comentarii indexate şi adnotate manual de către un utilizator. Conform datelor, Filtrul pentru lungimea comentariilor a obţinut valoarea cea mai bună pentru precizia medie calculată (Mean Average Precision [2]). Numărul total de comentarii relevante este de 50, însemnând că primele 10 rezultate pentru filtrul *Length* si filtrul *Entropy* au o rată de succes maximă.

Tabel 3 – Precizie medie pentru videoclipul Most Amazing Coincidence Ever (#nW3txVqGop4) din categoria Entertainment

Denumire filtru	Relevante găsite din 100	Precizie medie		
		10 rezultate	50 rezultate	100 rezultate
Length	40	0.2	0.43	0.56
Inform	26	0.08	0.16	0.23
Entropy	40	0.2	0.43	0.55
Wikipedia	29	0.17	0.27	0.34
Textalytics Sentiment	34	0.03	0.10	0.23
Textalytics Topics	30	0.03	0.11	0.20
Textalytics Classification	27	0.03	0.13	0.20
Textalytics SentimentPlus	32	0.01	0.08	0.19

Categorie: Music

Adnotarea videoclipului s-a efectuat pentru un număr de aproximativ 350 de comentarii indexate, dintre care 38 sunt considerate relevante de către utilizatori. Cel mai bun scor obţinut este prin aplicarea filtrului de relevanţă *Textalytics_Classification*.

[2] Mean Average Precision - https://www.kaggle.com/wiki/MeanAveragePrecision

Tabel 4 – Precizie medie pentru videoclipul Cut Chemist feat. Hymnal "What's the Altitude" Music Video (#7AVHXe-ol-s) din categoria Music

Denumire filtru	Relevante găsite din 100	Precizie medie		
		10 rezultate	50 rezultate	100 rezultate
Length	25	0.02	0.10	0.18
Inform	25	0.01	0.11	0.19
Entrophy	26	0.02	0.10	0.19
Wikipedia	15	0.00	0.02	0.05
Textalytics Sentiment	17	0.00	0.05	0.08
Textalytics Topics	23	0.04	0.10	0.16
Textalytics Classification	38	0.08	0.16	0.19
Textalytics SentimentPlus	21	0.00	0.05	0.11

Categorie: Education

Tabel 5 – Precizie medie pentru videoclipul 10 Unsolved Mysteries of the Internet (#9qHPqC1ZqZ4), din categoria Education

Denumire filtru	Relevante găsite din 100	Precizie medie		
		10 rezultate	50 rezultate	100 rezultate
Length	39	0.18	0.62	0.68
Inform	37	0.18	0.58	0.66
Entropy	39	0.16	0.58	0.66
Wikipedia	31	0.02	0.11	0.22
Textalytics Sentiment	29	0.00	0.12	0.19
Textalytics Topics	29	0.00	0.12	0.19
Textalytics Classification	23	0.00	0.05	0.11
Textalytics SentimentPlus	31	0.01	0.10	0.21

Categorie:People

Tabel 6 – Precizie medie pentru videoclipul Weather Versus Climate Change, (#cBdxDFpDp_k)

Denumire filtru	Relevante găsite	Precizie medie		
		10 rezultate	50 rezultate	100 rezultate
Length	64	0.12	0.48	0.70
Inform	55	0.11	0.41	0.56
Entropy	64	0.12	0.44	0.67
Badwords	37	0.06	0.12	0.21
Textalytics Sentiment	54	0.02	0.20	0.37
Textalytics Topics	50	0.01	0.06	0.25
Textalytics Classification	53	0.00	0.11	0.29
Textalytics SentimentPlus	53	0.04	0.17	0.35

Testele s-au efectuat folosind cinci filtre de căutare implementate: *Bad Words* – pentru identificarea cuvintelor obscene, negative, *Entropy* – promovează comentariile complexe din punct de vedere lexical, *Inform* – ține cont de unicitatea informației dintr-un comentariu, *Length* – ține cont de lungimea unui comentariu (numărul de cuvinte al acestuia), Textalytics Topics – promovează comentariile care au subiecte asemănătoare cu cele identificate pe baza titlului şi descrierii videoclipului, Textalytics Classification – promovează comentariile pentru care s-au identificat etichete similare cu cel obținute pentru titlu şi descriere, iar Textalytics Sentiment promovează comentariile în funcție de rezultatele analizei emoționale.

Concluzii şi cercetări viitoare

Proiectul combină un set de criterii pentru filtrarea comentariilor pentru a obține rezultate cât mai relevante. Comentariile sunt slabe în metadate şi din acest motiv este dificilă manipularea şi clasificarea lor. Din acest motiv este nevoie de o abordare dinamică, unde criteriile de relevanță sunt adaptabile la o serie întreagă de parametri, cum ar fi modelul semi-local al utilizatorului, factorii social şi adaptivitatea în funcție de interacțiunea în timp.

Pe lângă analiza datelor interne colectate la nivel de aplicație este necesară relaționarea datelor cu surse externe, maparea acestora cu concepte relaționate din domenii diferite, pentru a ne putea asigura că într-adevăr recomandările de comentarii pot fi utile pentru utilizator.

A fost efectuată o analiză a comentariilor pe baza unui grup heterogen de criterii şi factori pentru a obține un grad de relevanță cât mai apropiat de rezultatele obținute prin adnotarea manuală. În general, comentariile nu sunt documente bogate în metadate, iar pentru o obține rezultate cât mai relevante este utilă integrarea cu servicii externe de analiză a textului, ce pot fi folosite pentru obținerea de informații suplimentare. Totodată, utilizarea unor astfel de servicii este consumatoare de resurse şi de timp şi nu este fezabilă pentru analizarea unui număr foarte mare de comentarii atunci când există o constrângere de timp. Analiza relevanței comentariilor presupune testarea intensivă a criteriilor/atributelor de filtrare pentru un set cât mai mare şi cât mai variat de videoclipuri, iar în funcție de rezultatele

obținute, trebuie să se ajusteze automat sau manual ponderea fiecărui criteriu în calculul total al indexului de căutare.

De asemenea, pentru analiza preliminară a comentariilor, direct din modulul de indexare şi prelucare a datelor, pentru identificarea comentariilor de tip spam, se pot folosi tehnici deja studiate şi dezvoltate pentru analiza comentariilor spam publicate pe bloguri (Kamaliha, 2008).

Ținând cont de analiza efectuată la nivel de videoclip, rezultatele obținute sunt relevante, iar implementarea unei astfel de soluții în orice aplicație poate rezolva problema afişării comentariilor relevante în cazul în care numărul de comentarii publicate depăşeşte o anumită limită.

În concluzie, extragerea comentariilor relevante dintr-un set foarte mare de date se poate face folosind o gamă largă de criterii, iar performanța acestora este dependentă în mod direct de domeniu, categoria videoclipurilor şi profilul utilizatorilor. Ținând cont de aceste aspecte, este dificilă propunerea unei soluții general valabile pentru identificarea comentariilor relevante, însă performanța rezultatelor poate fi maximizată selectând criteriile potrivite contextului în care se face analiza.

Confirmare

Rezultatele prezentate în acest articol au fost parțial finanțate şi obținute cu sprijinul Ministerului Fondurilor Europene prin Programul Operațional Sectorial Dezvoltarea Resurselor Umane 2007-2013, Contract nr. POSDRU/159/1.5/S/132397.

Bibliografie

Serbanoiu, A. & Rebedea, T., 2013. Relevance-Based Ranking of Video Comments on YouTube*, Proceedings of CSCS 2013*, Bucharest.

Mishne, G., 2005. Blocking Blog Spam with Language Model Disagreement*, Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.

Cheng, X., 2007. Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study, *Procs of the 7th ACM SIGCOMM Conference on Internet Measurement*, San Diego (CA, USA).

Holpuch, A., 2013. Netflix and YouTube make up majority of US internet traffic, new report shows, *The Guardian.* [Online]
Available at: http://www.theguardian.com/technology/2013/nov/11/netflix-youtube-

dominate-us-internet-traffic
[Accessed 10 May 2014].

Chatzopoulou, G., Sheng, C. & Faloutsos, M., 2010. *A first step towards understanding popularity in YouTube,* Riverside: University of California.

Esuli, A. & Sebastiani, F., 2006. SENTIWORDNET: A Publicly Available Lexical Resource. *Proceedings of LREC. Vol. 6.*.

Han, Y.-S., Kim, L. & Cha, J.-W., 2009. *Evaluation of User Reputation,* Seoul: Online Communities and Social Computing.

Hsu, C.-F., Khabiri, E. & Caverlee, J., 2009. *Ranking Comments on the Social Web,* College Station,: Department of Computer Science and Engineering.

Junqueira, F. P. & Plachouras, V., 2007. *Workshop on Large-Scale Distributed Systems for Information Retrieval,* Barcelona: ACM SIGIR 2007.

Kamaliha, E., 2008. *Characterizing Network Motifs to Identify Spam Comments,* Pisa: IEEE.

Lange, P. G., 2007. *Commenting on Comments: Investigating Responses to Antagonism on YouTube,* Tampa: Annenberg Center for Communication .

Manning, C. D., Raghavan, P. & Schutze, H., 2008. *Introduction to Information Retrieval Vol. 1.* Cambridge: Cambridge university press.

Sierdorfer, S., Chelaru, S., Nejdl, W. & San Piedro, J., 2010. *How Useful are Your Comments?Analyzing and Predicting YouTube Comments and ratings,* Hannover: L3S Research Center.

Yue, Y., Finley, T., Radlinski, F. & Joachims, T., 2007. *A Support Vector Method for Optimizing Average Precision,* Seoul: ACM.

.