Universitatea Babeș-Bolyai Cluj-Napoca
Facultatea de Științe Economice și Gestiunea Afacerilor

Domeniu: Cibernetică și statistică

Metode Eficace pentru Analiza și Recomandarea Conținutului în Social Media

Rezumatul tezei de doctorat

Profesor Coordinator:
Prof. Dr. Tomai Nicolae

Candidat:
Muntean Cristina Ioana

-2012-

Cuprinsul tezei

1 Introducere 10
- 1.1 Motivație si strategie de cercetare 13
- 1.2 Problema de cercetare 14
- 1.3 Soluții propuse 15
- 1.4 Organizarea tezei 20

2 Social media și Twitter 21
- 2.1 Afacerile în contextul social 21
- 2.2 Descrierea și oportunitățile Social media 23
- 2.3 Twitter ca mediu pentru Social media 27
  - 2.3.1 Rețeaua socială Twitter 28
  - 2.3.2 Analiza rețelelor sociale 33
  - 2.3.3 Anatomia Twitter 35
- 2.4 Sumar 40

3 Angajarea utilizatorilor 42
- 3.1 Angajarea utilizatorilor 42
  - 3.1.1 Definirea angajării utilizatorilor 43
  - 3.1.2 Metrici de evaluare a angajării utilizatorilor 44
  - 3.1.3 Angajarea utilizatorilor în Social media 50
- 3.2 Gamification 53
  - 3.2.1 Definirea Gamification 57
  - 3.2.2 Mecanici și tehnici de joc 58
  - 3.2.3 Exemplu teoretic 61
  - 3.2.4 Gamification în Twitter 65
- 3.3 Sumar 69

4 Gestionarea Big Data 70
- 4.1 Colectarea de date 70
  - 4.1.1 Crawler-e web 71
  - 4.1.2 Streaming APIs 73
- 4.2 Tehnologii pentru procesarea Big Data 77
  - 4.2.1 Introducere în Big Data 78
  - 4.2.2 Apache Hadoop și MapReduce 80
  - 4.2.3 Cascading 84
  - 4.2.4 Apache Mahout 88
- 4.3 Sumar 93

5 Modele de analiză și recomandare de text 95

5.1 Sisteme de recomandări 96
5.2 Analiza de text 99
5.2.1 Datele textuale 100
5.2.2 Indexarea documentelor 101
5.2.3 Asignarea de greutăți vectorilor 103
5.3 Tehnici de învățare automată (Machine Learning) 105
5.3.1 Naïve Bayes 105
5.3.2 Arbori de decizie 109
5.3.3 Clusterizarea cu ajutorul K-Means 111
5.4 Information Retrieval și căutarea 114
5.4.1 Indexul inversat 114
5.4.2 Regăsirea documentelor 115
5.5 Tehnici de evaluare 116
5.5.1 Evaluarea rezultatelor neordonate 117
5.5.2 Evaluarea rezultatelor ordonate 119
5.6 Sumar 120

6 Articole similare și aplicații în Twitter 121
6.1 Probleme și aplicații 121
6.2 Clusterizarea în Twitter 126
6.3 Analiza și recomandarea de hashtag-uri 128
6.4 Sumar 130

7 Analiza, clusterizarea și recomandarea de hashtag-uri 132
7.1 Descrierea setului de date din Twitter 133
7.1.1 Statistici asupra setului de date 138
7.1.2 Utilitatea hashtag-urilor 140
7.2 Clusterizarea de hashtag-uri din Twitter 142
7.2.1 Setul de date 143
7.2.2 Descrierea și configurarea experimentelor 146
7.2.3 Rezultate 149
7.2.4 Remarci 151
7.3 Oferirea de recomandări de hashtag-uri din Twitter 153
7.3.1 Definirea problemei 155
7.3.2 Architectura sistemului 161
7.3.3 Experimente și rezultate 164
7.3.4 Remarci 168
7.4 Sumar 170

8 Concluzii 171
8.1 Concluzii și dezvoltări ulterioare 171

8.2 Diseminarea resultatelor 176
8.3 Lista completă a articolelor 177

Abstract

Expansiunea rețelelor sociale și cantitatea mare de date din Social Media stimulează apariția de aplicații interesante și tehnologii care să le suporte. Ne începem cercetarea prin a încerca să înțelegem Social Media și mecanismele ce determină interacțiunea utilizatorilor cu aceste platforme. Ca și studiu de caz am selectat Twitter, asupra căruia vom baza observațiile ulterioare, studiul și aplicațiile. Analizăm această platformă din două puncte de vedere: unul aparținând simplului utilizator și unul din perspectiva analizei rețelelor sociale. Dorim să abordăm prelucrarea datelor colectate din Twitter cu instrumentele adecvate, astfel inspectăm semnificația Big Data (colecții mari de date) și tehnologiile ce folosesc MapReduce pentru procesarea acestor date. Studiem modele de reprezentare a textului și oferirea de recomandări în acest sens, prin explorarea mai multor discipline: Information Retrieval (IR), Machine Learning (ML, învățare automată) și Sisteme de Recomandări. După consultarea articolelor similare pe tema cluterizării și oferirea de recomandări folosind conținut din Twitter, am propus propriile noastre modele. Am descărcat un set mare de date din Twitter pe o perioadă de trei săptămâni, set reprezentând 10% din tweet-urile ce apar în fluxul public de date. Am curățat și procesat datele. Asupra setului de date rezultat am facut o analiză aprofundată pentru a descoperi perspective variate și pentru a înțelege specificitatea problemei. Observăm că adesea hashtag-urile pot avea o formă criptică (abrevieri, acronime, argou sau cuvinte și numere concatenate). Dorim să explorăm sensul acestora folosind clusterizarea, adică gruparea hashtag-urilor în clustere. Observăm că termenii cei mai frecvenți din clusterele formate reușesc să traducă sensul unor hashtag-uri puțin evidente sau mai greu de înțeles și să le încadreze în context. O altă aplicație pe care o propunem este un sistem de recomandare de hashtag-uri. După analizarea tiparelor de etichetare a tweet-urilor cu ajutorul hashtag-urilor reușim să descoperim natura lor dual\u0103, primul tipar constând în înlocuirea cuvintelor sau frazelor cu un hashtag conținând acel cuvânt sau acea frază (hashtag inline), iar al doilea tipar oferă contextul sau o categorie informală căreia îi poate fi atribuit tweet-ul în funcție de semantica acestuia (hashtag de context). Propunem un model ce îmbină ambele tipuri de comportamente și tratăm modelul de recomandare pentru hashtag-uri inline ca o problemă de predicție, iar cel pentru hashtag-uri contextuale ca o problemă de IR. Sistemul propus oferă rezultate mai bune decât alternativa propusă de literatură. Totodată propunem soluții privind îmbunătățirea rezultatelor.

Cuvinte cheie: Social Media, Gamification, Big Data, Twitter, clusterizare, sisteme de recomandări.

Cuprins

1 Introducere **6**
1.1 Motivație 6
1.2 Obiective 7
1.3 Organizarea tezei 8

2 Social media și Twitter **10**

3 Angajarea utilizatorilor și Gamification **12**

4 Gestionarea Big Data **14**

5 Modele de analiză și recomandare de text **16**

6 Articole similare și aplicații în Twitter **18**

7 Analiza, clusterizarea și recomandarea de hashtag-uri **20**

8 Concluzii **26**
8.1 Concluzii și direcții viitoare 26
8.2 Diseminarea rezultatelor din teză 29
8.3 Lista completă de articole 30

Capitolul 1

Introducere

Atât utilizatorii cât și companiile ce activează online trebuie să înțeleagă puterea de care dispun și să profite de opotunitățile ce pot să apară din marea masă de informații ce îi întâmpină în acest mediu. Datorită faptului că utilizarea Internetului a crescut exponențial în ultimii ani, ne pune în postura de a filtra, sorta și selecta informațiile ce ne pot fi utile dintr-o mare de date nestructurate și instabile. Acest lucru nu transpare utilizatorilor simpli de Internet, însă companii precum Google, Twitter, Facebook, Amazon, Last.fm, se confruntă cu aceste probleme în oferirea serviciilor lor. Acestea încearcă să optimizeze stocarea datelor, transformarea informațiilor în cunoștiințe și oferirea de produse, servicii relevante și personalizate către utilizatori, pentru a le maximiza satisfacția.

Scopul principal al tezei este descoperirea de metode prin care se pot îmbunătăți și stimula interacțiunile între utilizatori și uneltele, serviciile, platformele disponibile pe Internet.

1.1 Motivație

Faptul că utilizatorii de Internet beneficiază de o multitudine de facilități ne permite să descoperim metode diverse de a construi altele noi sau de a îmbunătăți cele existente. Motivația noastră constă în a găsi noi moduri de a folosi sau exploata informația, de dimensiuni și complexitate mare, pentru a ușura utilizarea acesteia și pentru a crea unelte noi ce pot ajuta actorii ce activează online. Datele pe Internet evoluează și se schimbă rapid, iar cercetarea trebuie să țină pasul.

Cercetarea pe care o conducem are un caracter interdisciplinar. Folosim metode de cercetare ce iau în calcul dimensiunile variate ale problemei. În același timp oferim soluții pragmatice pentru obiectivele setate, și anume

analiza textului scurt, clusterizare și oferirea de recomandări pentru un mediu aparte ce facilitează Social Media, Twitter.

1.2 Obiective

Datorită faptului că Social Media a apărut destul de recent, ne confruntăm cu dezvoltări și schimbări rapide în acest domeniu. Astfel că provocările și problemele ce derivă din Social Media s-au schimbat în timp, însă în studiul nostru încercăm să ținem pasul cu acest subiect. Aplicațiile ce le propunem iau în considerare teme noi și interesante ce exploatează particularități de structură și conținut.

Principalul obiectiv ale acestei cercetări este găsirea de metode eficiente pentru analiza Social Media și propunerea de aplicații utile pentru utilizatorii acesteia, astfel încât să beneficieze de cele mai bune facilități și informații din acest mediu.

Pentru abordarea subiectului ne-am setat următoarele *obiective specifice*:

1. Dorim să facem o analiză aprofundată a Social Media și a rețelelor sociale ce oferă suportul pentru acest tip de schimb de informații.

2. Dorim să studiem și să înțelegem de ce platformele ce facilitează Social Media reușesc să atragă utilizatorii, care sunt avantajele și dezavantajele de a deveni o componentă activă în astfel de comunități online și cum pot fi utilizatorii stimulați să interacționeze cu acest tip de activități, anume împărtășirea de informații în contextul social, și platforme.

3. Dorim să studiem particularitățile de conținut ce pot fi găsite în aceste medii și să descoperim aplicații relevante ce ar putea îmbunătății experiența utilizatorilor și analiza datelor. Vrem să realizăm cele propuse prin:

   (a) Studiul literaturii în domeniu și a ultimelor articole în ceea ce privește aplicațiile și analiza platformelor ce suportă Social Media și microblogging.

   (b) Colectarea și extragerea de date disponibile din platformele de Social Media.

   (c) Folosirea de uneltele adecvate pentru analiza și modelarea cantităților mari de date disponibile în astfel de medii.

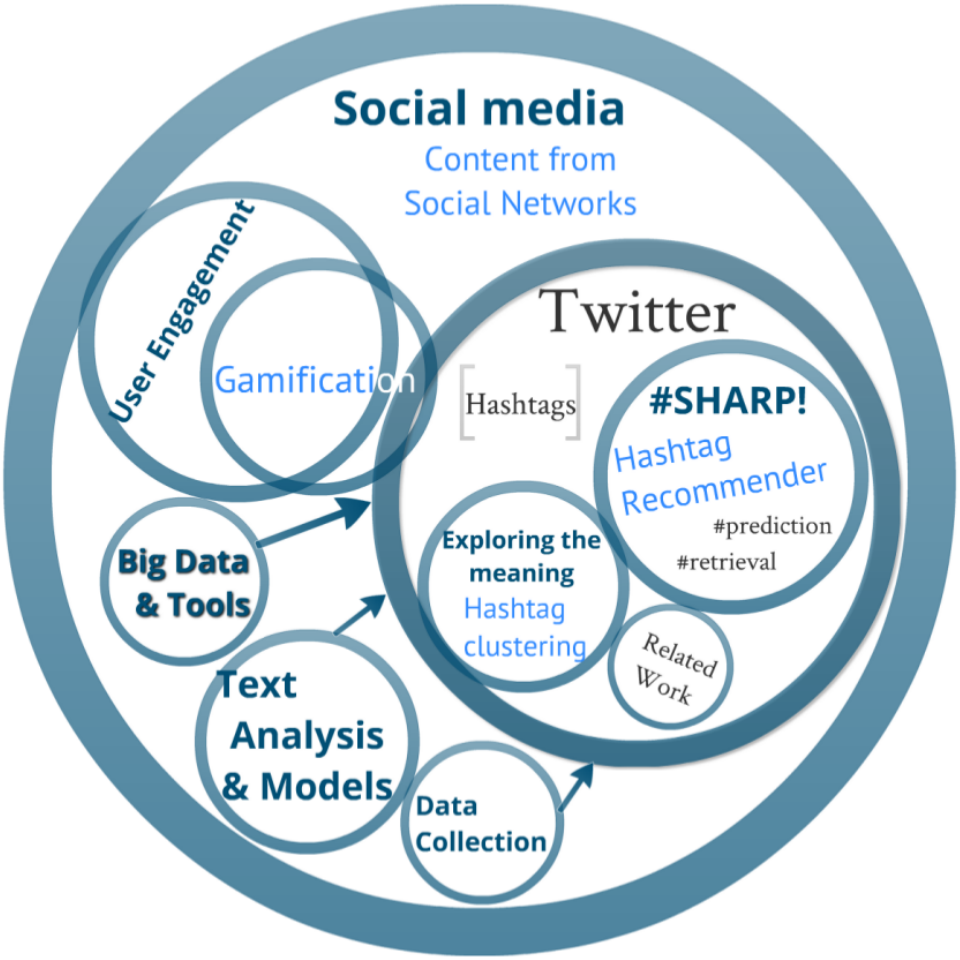

Figura 1.1: Organizarea tezei

(d) Considerarea algoritmilor și metodologiilor potrivite pentru analiza textului, precum metode de învățare supervizate și non-supervizate pentru minarea conținutului.

(e) Propunerea de aplicații proprii și confruntarea acestora cu lucrări reprezentând cele mai recente cercetări în domeniu.

1.3 Organizarea tezei

Având în vedere obiectivele propuse anterior, ne concentrăm atenția pe posibilele moduri de abordare a obiectivelor și oferirea de soluții posibile. O vedere de ansamblu asupra tezei poate fi văzută în Figura 1.1.

Parte din cercetarea prezentată în această teză a fost realizată pe par-

cursul unui stagiu de 10 luni la laboratorul High Performance Computing, in cadrul ISTI CNR, Pisa, sub supravegherea Dr. Fabrizio Silvestri și împreună cu echipele de cercetare menționate în [12, 7, 6, 25].

În Capitolul 2 prezentăm contextul social pe Internet și cum acesta a evoluat în ultimii ani. Descriem Social Media, identificăm cele mai importante canale și exemplificăm conceptele folosind ca exponent Twitter. Analizăm platformele și serviciile disponibile și explicăm cum pot fi folosite la capacitatea lor maximă. În Capitolul 3 descriem elementele care îi fac pe utilizatori să activeze pe platformele sociale și analizăm experiența utilizatorilor și stimularea angajării acestora în varii activități. O posibilă modalitate o reprezintă Gamification. Explicăm în detaliu ceea ce înseamnă, impactul pe care îl are, cum Twitter este analizat prin prisma Gamification și propunem un model prin care se aplică elemente de Gamification unei aplicații sociale de e-learning. În Capitolul 4 explicăm cum se pot colecta date de pe Internet în general, iar din Twitter în particular și analizăm tehnologiile ce permit procesarea cantităților mari de date și modelarea datelor în mod distribut. Aplicăm aceste instrumente și tehnologii la datele colectate din Twitter datorită mărimii setului de date. În Capitolul 5 studiem modurile posibile în care se pot modela sau recomanda date textuale și descriem metodologiile pe care le-am folosit și le-am implementat în cercetarea noastră. Deasemenea discutămi modalități și metrici de evaluare. În Capitolul 6 facem un studiu detaliat despre cum textul scurt din Twitter este abordat în cercetare și prezentăm și cele mai importante aplicații propuse în literatură. De asemenea analizăm și cele mai recente cercetări în ceea ce privește clusterizarea și recomandarea de text în Twitter. În Capitolul 7 prezentăm statistici despre setul de date pe care l-am extras din Twitter și propunem propriul mod de abordare a clusterizării și recomandării de hashtag-uri din Twitter, motivăm utilitatea acestora precum și prezentăm rezultatele obținute. În Capitolul 8 prezentăm concluziile și direcțiile ulterioare de cercetare.

Capitolul 2

Social media și Twitter

În acest capitol începem prin a introduce conceptele principale ale cercetării. Descriem pe scurt interacțiunile dintre utilizatori, clienți și companii în contextual web-ului social și a conținutului generat de utilizatori. Interacțiunile utilizatorilor înseamnă creare de valoare pentru companiile de pe Internet, de aceea este în interesul acestora să stimuleze angajarea utilizatorilor și activitatea acestora.

Tapscott și Williams [43], în cartea lor „Wikinomics: How Mass Collaboration Changes Everything", elaborează asupra noului univers economic creat pe Internet, unde valorile culturale precum participarea, colectivismul și creativitatea sunt promovate. Câteva exemple de astfel de inițiative sunt YouTube, MySpace, Wikipedia, Flickr, Second Life, Linux sau Twitter. Ele sunt create și alimentate de grupuri anonime de utilizatori ce se exprimă în propriul mod, cu mult mai puține constângeri decât în modelul Web 1.0. Astfel, în locul modelului de business ierarhic de la producător la consumator, întâlnim modelul *co-creației* [41]. Prin creativitatea de masă, producția de la egal la egal (peer production) și co-creație, distanța între producția colectivă (non-mercantil, public) și comercială (mercantil, privat) devine mai mică, precum și distanța dintre producători și consumatori. Studiem mecanismele ce susțin Social Media: co-creația, marketingul prin conversație sau inovația deschisă, și comportamentul utilizatorilor în acest mediu. Potrivit lui Osterwalder [33, 34, 35], Social Media apare ca suma platformelor (bloguri, wiki-uri, rețele sociale sau alte platforme unde utilizatorul generează conținutul) ce permit și încurajează interacțiunea socială prin partajarea de conținut și discuții.

În [22] autorii susțin că Social Media reprezintă o unealtă eficientă, iar companiile ce ezită sau nu sunt în stare să înțeleagă sau să dezvolte strategii Social Media, pierd oportunități și în același timp sunt amenințate de

competiția care știe să folosească acest avantaj. Autorii identifică principalele *functionalități* ale Social Media: identitate, conversații, partajare, prezență, relații, reputație și grupuri. Li și Bernoff în [24] împart *participanții* la Social Media în: creatori, colecționari, utilizatori ce crează legături și simpli spectatori.

Precum am menționat anterior, Social Media are numeroși facilitatori, însă marea parte a activității este susținută de rețele sociale, astfel că direcționăm discuția către Twitter, studiul nostru de caz.

Twitter este o rețea socială și o platformă de microblogging cu relații într-un singur sens între utilizatori, platformă ce permite postarea de scurte mesaje ca status, conținând URL-uri, mențiuni de alți utilizatori și/sau hashtags. Twitter este de asemenea privit ca o *rețea de informare in timp real*. Analizăm Twitter dintr-o dublă perspectivă, aceea a utilizatorului și din punctul de vedere al analizei rețelelor sociale, menționând cele mai relevante studii ce tratează această platformă ca o rețea socială.

O rețea socială este un model de prietenie, consiliere, comunicare sau suport ce există între membrii unui sistem social [23, 11]. Twitter se autointitulează: „Cea mai rapidă și simplă cale de a sta aproape de tot ceea ce te interesează." Unește utilizatorul cu ultimele povestiri, știri, idei și opinii, tot ceea ce trebuie să facă utilizatorul este să găsească acele conturi care îi sunt de interes și să le urmărească. Oferim de asemenea o trecere în revistă a literaturii despre perspectiva analizei rețelelor sociale aplicată pe Twitter. S-a descoperit că în ciuda relațiilor uni-direcționale dintre utilizatori, gradul de separare dintre aceștia este aproximativ 4, utilizatorii urmează o distribuție exponențială (power law[1]) și există un grad scăzut de reciprocitate între aceștia. Utilizatorii au intenții diverse: căutarea de informații, împărtășirea de informații sau relații de prietenie. Putem concluziona că Twitter nu este o rețea socială tipică, fapt ce face studiul acesteia interesant, motivându-ne să analizăm acest mediu și în capitolele următoare.

[1]http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html

Capitolul 3

Angajarea utilizatorilor și Gamification

Precum am menționat anterior, pentru companiile ce activează pe Internet, interacțiunile din parte utilizatorilor (navigare, recomandare, comentare și împartășirea de conținut) contribuie la creșterea valorii acestora. Este importantă crearea unei baze de clienți și stimularea acestora în a se reîntoarce la website sau să impărtășească conținutul acestuia cu rețeaua lor socială. În acest capitol studiem cum pot fi utilizatorii unui website stimulați să activeze, care sunt caracteristicile și metricile pentru măsurarea implicării acestora (User Engagement).

Potrivit [3] angajarea utilizatorilor este „conexiunea între emoțional, cognitiv și/sau comportamental, ce există în orice moment în timp și de-a lungul timpului, între un utilizator și o resursă tehnologică". Caracteristicile sunt următoarele: atenție direcționată, stimul/emoție pozitivă, estetică, anduranță, noutate, control, reputație, încredere și așteptări, motivație, stimulente și beneficii. În [38] autorii propun 8 *indici* pentru măsurarea implicării utilizatorilor: indexul de adâncime a click-urilor, indexul de durată, indexul de regență, indexul de loialitate, indexul de brand, indexul de feedback (răspuns) și indexul de interacțiune.

În contextual social nu este de ajuns să măsurăm doar interacțiunea unui utilizator cu o bucată de conținut, este necesar să luăm în considerare relațiile sociale și interacțiunile dintre utilizatori.

Dacă anterior am argumentat importanța angajării și stimulării utilizatorilor, ne propunem să oferim și o soluție pentru acesta, și anume prin Gamification. Gamification [17] este „utilizarea mecanicilor de joc în aplicații ce nu reprezintă jocuri". Orice aplicație, atribuție sau context poate fi teoretic gamificat. Scopul principal este de a stimula angajamentul utilizatorilor folosind tehnici de joc precum clasamente cu punctaje sau feedback perso-

nalizat rapid [19], astfel încât aceștia capătă un sens al apartenenței și un sens atunci când interacționează cu o sarcină [36]. Principalele moduri de a crește implicarea utilizatorilor prin gamification sunt: cicluri de feedback accelerate, reguli și obiective bine definite, narațiune convingătoare și sarcini ce pot fi obținute. Explicăm de asemenea mecanismele din spatele jocului și a angajării utilizatorilor. Mecanicile de joc sunt caracterizate de următoarele 3 atribute:

- Tipul de mecanică de joc: progresie, feedback și comportamental;
- Beneficii: implicare/angajare, loialitate, timp petrecut, influență, amuzament, SEO, conținut generat de utilizatori, viralitate;
- Tipuri de personalitate [8]: exploratori, realizatori, prieteni sau criminali;

În acest capitol discutăm de asemenea diferite mecanici de joc și tehnici de Gamification și dăm un exemplu de gamificare a unei aplicații (folosind punctaje, clasamente, statuturi, nivele, etc.). Am analizat și Twitter prin prisma Gamification. Observăm câteva *statistici de angajare* si *cicluri de feedback* ce sugerează Gamification: numărul de urmăritori, retweet-uri, mesaje directe, mențiuni (@), taguri (#) și liste. Chiar dacă nu există un clasament formal și global platformei, utilizatorii sunt conștienți de propriul lor status, reflectat de elementele mai sus menționate. Datorită faptului că profilele sunt publice, utilizatorii sunt motivați și se implică în ciclul de feedback al platformei. Este notabil că Gamification se află la baza acestui serviciu, chiar dacă nu este evident.

Capitolul 4

Gestionarea Big Data

În acest capitol expunem o abordare pragmatică asupra datelor si uneltelor folosite, concentrându-ne pe explicarea tehnologiilor folosite pentru colectarea datelor din Social Media sau website-uri și pe instrumentele folosite în procesarea cantităților mari de date. Astfel am implementat două metode de colectare a datelor, cu ajutorul *crawler*-elor web și a *streaming API*. Prezentăm implementările realizate și seturile de date colectate.

Folosim un crawler web pentru a naviga în web sau în interiorul unui website și a colectata conținutul paginilor. Am creat un astfel de crawler pentru a extrage utilizatori din eBay[1] împreună cu feedback-ul (positive, negative, neutral) corespondent primit de la alți utilizatori. Un al doilea crawler l-am creat pentru a extrage postări de anunțuri de slujbe pe site-url BestJobs[2]. Ambele crawlere au fost implementate în Python iar fiecare are propria specificitate.

Pentru a extrage date din Twitter folosim Twitter Streaming API. La cererea noastră am primit de la Twitter nivelul de acces Gardenhose, destinat pentru scopuri de cercetare, ce oferă 10% din fluxul public de date (nivelul standard de accesibilitate, Sprizer, oferă doar 1%). Rata medie de download a fost de aproximativ 14.700 tweet-uri pe minut. Astfel am colectat un set de date de dimeniuni mari pe parcursul a trei săptămâni, constând în 443,288,820 tweet-uri în total, aproximativ 200 G de date comprimate, dimensiunea crescând de 10 ori în formatul necomprimat. Tweet-urile sunt livrate în format JSON.

Pentru a putea stoca și procesa această cantitate mare de date, explorăm tehnologiile specializate pe Big Data. Potrivit lui Edd Dumbill de la O'Reilly Radar [18], Big Data reprezintă datele ce depășesc capacitatea de procesare

[1] http://www.ebay.com/

[2] http://www.bestjobs.ro/

a bazelor de date convenționale.

Calculele în memorie au un cost ridicat când vine vorba de procesarea cantităților mari de date. Potrivit [45] există 2 factori determinanți ce au contribuit la dezvoltarea acestor tehnologii: creșterea masivă a rețelelor sociale și dezvoltarea MapReduce [16]. Bazele de date relaționale sunt depășite de acest tip de date.

Printre aceste tehnologii prezentăm Apache Hadoop, Cascading și Apache Mahout, toate arhitecturii ce permit calculul paralel și pe care le-am folosit și în procesarea datelor noastre, cum este explicat și în Capitolul 7. Cu ajutorul acestor platforme și librării am procesat datele urmând paradigma MapReduce. Apache Hadoop oferă un nivel de abstractizare asupra unui cluster format din mai multe mașini (calculatoare). De asemenea am stocat datele în mod distribuit pe HDFS[3].

Cascading este o librărie open source în Java ce permite procesarea de date în MapReduce construită deasupra Hadoop. Din dataset-ul extras din Twitter am eliminat informațiile nefolositoare și am rulat mai multe procese Cascading prin care am curățat datatele de tweet-uri în alte limbi decât în engleză, tweet-uri fără hashtag-uri, tweet-uri ce conțin doar URL-uri și hashtag-uri (considerate spam), etc. Intrările pentru fiecare operație provin de la o sursă (`Source`), sunt prelucrate în unul sau mai multe canale (`Pipes`) ce permit funcții sau proceduri variate (ex. Each, Every, GroupBy, CoGroup, SubAssembly) și ulterior sunt livrate în fișierele de ieșire `Sinks`.

Apache Mahout este o librărie MapReduce ce implementează algoritmi de Machine Learning (învățare automată). Este compusă din 3 module principale: algoritmi de recomandare, algoritmi de clasificare și algoritmi de clusterizare. Descriem mai detaliat secțiunea despre clusterizare, explicăm implementarea Mahout a K-means, algoritm folosit pentru clusterizarea hashtag-urilor din Twitter [31]. Bazele dezvoltării acestei platforme se regăsesc în [14], unde autorii explică cum diverși algoritmi de învățare pot fi paralelizați.

[3]Hadoop Distributed File System, adică sistemul distribuit de fișiere Hadoop.

Capitolul 5

Modele de analiză și recomandare de text

După ce am explicat care sunt instrumentele folosite pentru manipularea a mari cantități de date, în acest capitol explorăm modele de analiză și recomandare de text. Deoarece obiectivul nostru a fost să analizăm conținutul din Twitter, să clusterizăm hashtag-uri și să oferim recomandări de hashtag-uri, dorim să explicăm modelele fundamentale folosite pentru reprezentarea colecției de tweet-uri și hashtag-uri. Descriem atât structurile de date cât și algoritmii folosiți.

Meta-problema în această parte a cercetării este setarea bazelor pentru analiza și recomandarea de text din Twitter. Abordăm această problemă din perspective variate, deoarece granițele între abordări se întrepătrund.

Utilizatorii sunt copleșiți de *surplusul de interacțiune* și *surplusul de informație* existent pe Internet și în rețelele sociale. Pentru a contracara aceste probleme sunt propuse varii metode din domenii precum *Information Retrieval*, *Information Filtering* și *Sisteme de recomandări*. Explicăm principiile de bază din analiza de text și tehnici NLP[1], de la selectarea proprietăților textului, independența pozițională și condițională, vectori și scheme de cântărire precum *tfidf* și *BM25*, la indexarea documentelor și indexul inversat. Facem o trecere în revistă a algoritmilor de Machine Learning folosiți în experimentele ulterioare.

Reprezentăm hashtag-urile ca și "bag-of-words", deci orice hashtag este compus din toate tweet-urile în care el apare. Am selectat proprietățile textului (cuvinte semnificative) ce sunt relevante și am eliminat cele ce nu sunt, precum cuvinte foarte frecvente, și am unificat forma acestora prin

[1]Natural Language Processing, adică procesarea limbajului natural

păstrarea rădăcinii cuvintelor. Am folosit VSM[2] și $tfidf$ pentru a reprezenta fiecare hashtag ca și vector.

In ultima secțiune a capitolului discutăm metodele și metricile de evaluare folosite pentru validarea rezultatelor. Considerăm atât metrici de evaluare pentru rezultate a căror ordine contează, cât și metrici pentru rezultate a căror ordine nu contează. Astfel explicăm pe scurt precizia, recall-ul, F-measure, NDCG, P@k, MAP.

[2]Vector Space Model

Capitolul 6

Articole similare și aplicații în Twitter

În Capitolul 6 prezentăm cercetări similare și alte aplicații privitoare la Twitter prezente în literatură: difuziunea informațiilor [49, 4], detectarea trend-urilor [26, 2], analiza sentimentelor [15, 21, 1], detectarea de spam [44, 46, 9] sau recomandarea de știri în timp real [40, 42].

În ultima parte a capitolului ne referim cu precădere la aplicațiile propuse în Capitolul 7. Având în vedere obiectivul de a clusteriza hashtag-uri, am restrâns discuția la clusterizarea de text și text scurt. În ceea ce privește aplicația de clustering de text scurt, în [5] autorii clusterizează text din feed-ul RSS/Atom prin îmbogățirea puținelor informații disponibile în feed adăugând text adițional din Wikipedia. Articole despre tehnici de Machine Learning aplicate în Twitter abordează subiecte precum sumarizare sau detectare de subiecte prin LDA [32], clusterizare [48] și dezambiguizare a subiectelor sau clasificare [39, 37].

Majoritatea studiilor privitoare la clusterizarea în Twitter includ algoritmi de modelare a subiectelor (topic modeling). În [39] autorii folosesc LDA pentru a clasifica texte scurte și texte non-dense folosind seturi de date de mărimi considerabile. Sistemele de recomandări folosesc clusterizarea ca un pas premergător oferirii de sugestii. În [13] autorii sugerează tweet-uri pe baza istoricului unui utilizator și modelarea subiectelor. Transformă textul în vectori folosind VSM și $tfidf$. În mod similar, în TwitterRank [47] autorii încearcă să găsească utilizatorii inluenți din Twitter creând modele individuale pentru fiecare utilizator cu ajutorul LDA.

În [20] autorii analizează utilizarea hashtag-urilor pentru etichetarea tweet-urilor și compară comportamentele de etichetare dintre utilizatorii Twitter

și utilizatorii Delicious[1]. În [10] autorii încearcă să detecteze subiecte importante luând în considerare structura rețelei.

În ultima parte prezentăm un articol similar cu aplicația propusă de noi, recomandarea de hashtag-uri. După studiul nostru, putem spune că [50] de Zangerle *et al.* este singurul articol ce mai investighează recomandările de hashtag-uri. În acest articol autorii exploatează similaritatea între tweet-ul curent și tweet-uri mai vechi pentru a genera recomandări de hashtag-uri pe baza activității anterioare a utilizatorilor. Tweet-urile cele mai similare cu cel curent sunt extrase din setul de date pe baza similarității $tfidf$. Hashtag-urile rezultate din acest set de tweet-uri sunt extrase și ordonate folosind trei metode:

Overall Popularity Rank - ordonează candidații în funcție de popularitatea lor globală în setul de date.

Recommendation Popularity Rank - ordonează candidații in funcție de numărul de apariții în setul intermediar extras.

Similarity Rank - calculează similaritatea între două tweet-uri folosind $tfidf$, producând astfel candidații pentru recomandări.

Rezultatele sunt evaluate folosind precizie și recall, iar metoda cea mai eficientă este *Similarity Rank*.

[1] http://delicious.com/

Capitolul 7

Analiza, clusterizarea și recomandarea de hashtag-uri

În acest capitol ne prezentăm propriul dataset și modelele propuse, rezultatele fazei de experimentare și evaluarea acestora. Dacă în Capitolul 4 am prezentat modul de colectare al datelor, aici descriem datasetul obținut și o parte din pașii de preprocesare abordați. Din întregul set de date am eliminat tweet-urile ce nu sunt în engleză, retweet-urile, tweet-urile conversaționale si tweet-urile ce conțin spam. Prezentăm statisticile ce descriu setul de date obținut după preprocesare. Observăm că hashtag-urile urmează o distribuție *power law*, precum se poate observa în Figura 7.1, ceea ce înseamnă că puține hashtag-uri sunt extrem de frecvente în timp ce majoritatea se repetă doar de puține ori. Acest lucru se datorează faptului că hashtag-urile sunt generate de către utilizatori și nu există restricții în privința creării lor. Pot consta în acronime, greșeli de scriere, etc.; nu există nici un pas pentru validarea lor în momentul în care sunt lansate în rețea. Dataset-ul este descris în Tabelul 7.1.

Setul de date Twitter – (2 Decembrie – 23 Decembrie)	
Zile de activitate	21
Tweet-uri pe zi	21,108,991
Tweet-uri pe zi (în engleză) - cu cel puțin un hashtag - fără retweet-uri sau mențiuni	9,014,780 1,576,905 987,892
Hashtag-uri distincte pe zi	283,915

Tabela 7.1: Proprietățile (în medie) a setului de date din Twitter.

Pentru ca clusterizarea și recomandarea de hashtag-uri să aibă sens, mai

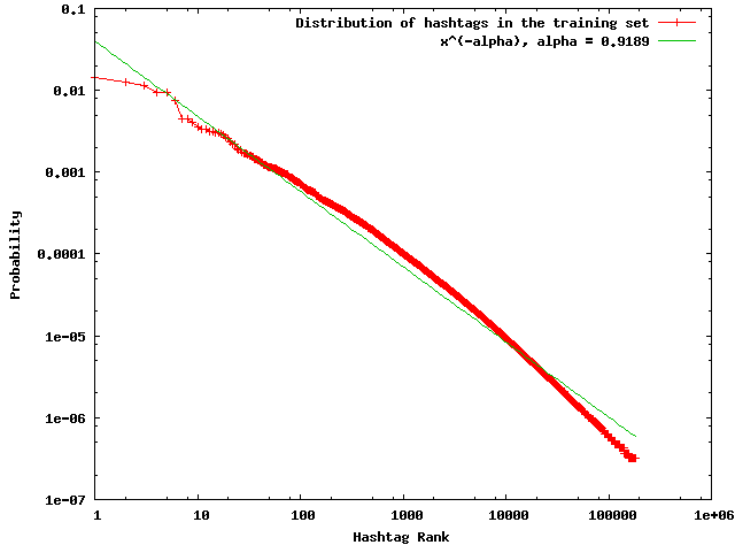

Figura 7.1: Distribuția de probabilitate a hashtag-urilor în setul de antrenament.

întâi dorim să dovedim utilitatea hashtag-urilor. Hashtag-urile măresc vizibilitatea unui tweet deasuprea nivelului de prietenie, deci prin folosirea hashtag-urilor în tweet-uri, acestea pot deveni vizibile în întreaga rețea pentru cei interesați de subiectul ce-l sintetizează; hashtag-urile ajută la gruparea și structurarea a întregi conversații. Observăm că probabilitatea ca un tweet să devină retweet[1] este semnificativ mai mare când hashtag-urile sunt folosite, rezultatele pot fi observate în Tabelul 7.2.

Probabilitate	Valoare
$P(R\|H)$	0.25%
$P(R\|\overline{H})$	0.16%

Tabela 7.2: Probabilitatea condițională a unui retweet pentru un tweet t: i) ce conține hashtag-uri $t \in H$, ii) ce nu conține hashtag-uri $t \in \overline{H}$.

Prima aplicație descrisă este clusterizarea hashtag-urilor, ce conform cunoștințelor noastre în literatură nu s-a mai făcut. Motivația în spatele acestei aplicații este simplă, ea putând fi utilizată în numeroase contexte. Precum am menționat anterior hashtag-urile prezintă o multitudine de particularități,

[1]Retweet-ul este un mecanism de broadcast ce consolidează și confirmă calitatea unui tweet. Un utilizator ce observă un tweet poate să facă retweet la acesta, astfel transmițându-l mai departe la propriul set de prieteni (urmăritori).

pot deveni chiar criptice. Ele sunt formate din acronime, cuvinte concatenate, cuvinte și numere concatenate, practic este dificil să fie interpretate în mod automatic. Ideea noastră este de a clusteriza hashtag-uri pe baza tweet-urilor în care apar. Pentru fiecare hashtag am creat un așa numit document virtual, ce funcționează ca un dicționar pentru sensul acelui termen. Dorim să capturăm sensul lor semantic prin similaritatea cu alte hashtag-uri mai explicite sau din termenii cei mai frecvenți din documentele virtuale clusterizate pe baza similarității. În experimente folosim seturi de date zilnice, ce le transformăm în asocieri `<hashtag, document vitual>`, pe care le convertim în vectori și le clusterizăm folosind K-means. Experimentăm cu un număr variabil de clustere, k=$\{20, 40, 80, 100, 200, \ldots, 1000\}$. Rezultatele sunt încurajatoare. Dacă pentru un număr mic de clustere, grupurile nu sunt foarte clare, pentru un număr mai mare de clustere, acestea sunt mult mai bine delimitate, iar subiectul corespunzător grupului este ușor de identificat. Prezentăm cele mai frecvente hashtag-uri dintr-un cluster împreună cu cei mai frecvenți termeni, iar corespondența este clară, cum se poate observa în Tabelul 7.3. Calculăm distanța inter și intra clustere în Figura 7.2, iar rezultatele arată că pentru $k < 100$ clusterele nu sunt bine delimitate, pe când pentru $k \geq 100$ sunt mai bine definite. Folosim dimensiunea lui k pentru a seta precizia și granularitatea subiectelor pe care dorim să le descoperim. Pentru un k mare subiectele sunt mai specifice, iar pentru un k mai mic aceste sunt mai generale.

Termeni frecvenți	occupy, ows, wall, street, protest, ndaa, movement, afghanistan, noccupy, st.
Hashtag-uri frecvente	ndaa, ows, occupy, occupywallstreet, china, peace, yyc, economy, kpop, washington.

Tabela 7.3: Exemplu de cluster pentru Dataset15 cu $k = 500$

Cea de-a doua aplicație este *recomandarea de hashtag-uri*. Pentru această aplicație am construit un sistem numit #SHARP! , precum descris în Figura 7.3. Am analizat hashtag-urile și am descoperit o diviziune a acestora în funcție de scop și funcționalitate. Astfel că hashtag-urile sunt împărțite în inline (hashtag-uri ce înlocuiesc cuvinte dintr-un tweet) și contextuale (hashtag-uri ce caraterizează subiectul unui tweet iar termenul/termenii acestuia nu sunt incluse în tweet). În funcție de cele două tipologii, tratăm problema recomandărilor dintr-o dublă perspectivă: considerăm recomandările inline ca o problemă de predicție, iar cele contextuale ca o problemă de Information Retrieval. Reunim ambele seturi de recomandări într-o listă unică, combinăm rezultatele și le prezentăm utilizatorului final ca lista finală de recomandări.

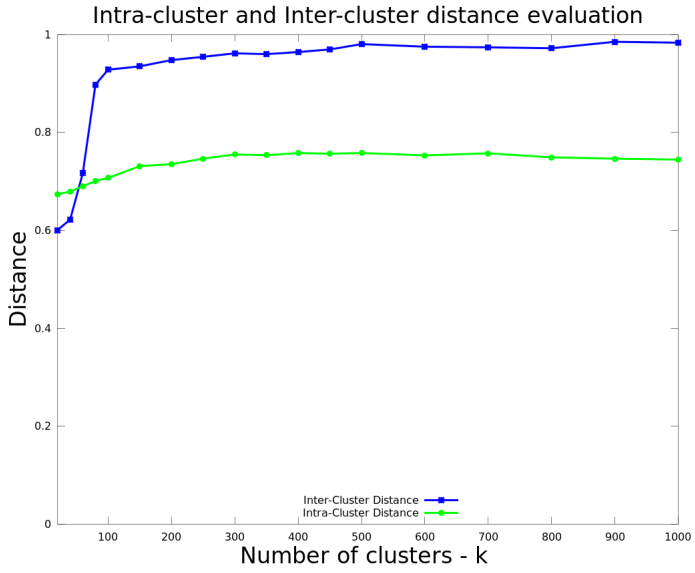

Figura 7.2: Evaluarea K-means pentru Dataset14 variind k

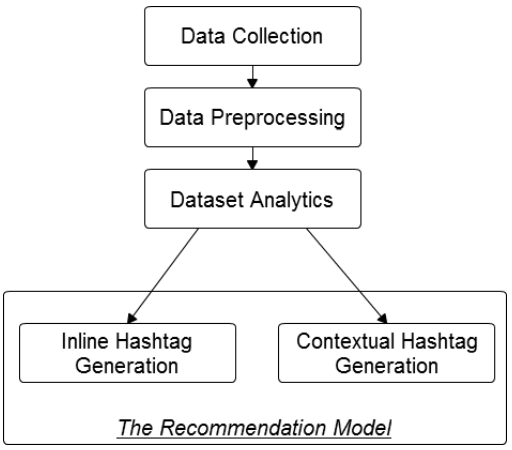

Figura 7.3: Arhitectura #SHARP!

	Measure	*Clairv.*	*Zangerle*	#SHARP!				
				$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
	Precision	0.0001	0.116	0.107	**0.125**	**0.121**	0.092	0.085
	Recall	0.0001	0.092	0.085	**0.099**	**0.094**	0.069	0.064
@1	F1	0.0001	0.099	0.091	**0.106**	**0.101**	0.075	0.069
	F2	0.0001	0.094	0.087	**0.101**	**0.096**	0.071	0.065
	NDCG	0.0001	0.116	0.107	**0.125**	**0.121**	0.092	0.085
	Precision	0.002	0.078	0.063	0.073	0.075	0.069	0.057
	Recall	0.004	0.176	0.143	0.162	0.167	0.153	0.127
@3	F1	0.003	0.103	0.083	0.095	0.098	0.090	0.075
	F2	0.003	0.135	0.109	0.124	0.128	0.118	0.097
	NDCG	0.001	0.092	0.073	0.084	0.085	0.075	0.063
	Precision	0.002	0.062	0.047	0.052	0.054	0.053	0.046
	Recall	0.009	0.228	0.173	0.191	0.198	0.195	0.168
@5	F1	0.003	0.093	0.070	0.078	0.081	0.079	0.069
	F2	0.005	0.141	0.106	0.118	0.122	0.120	0.103
	NDCG	0.002	0.081	0.059	0.067	0.068	0.062	0.054

Tabela 7.4: Performanta sistemului #SHARP! în comparație cu *Clairvoyant* și *Zangerle*.

Descriem în detaliu cum recomandările sunt modelate și cum sunt realizate experimentele.

Rezultatele sunt comparate cu 2 alte sisteme, *Clairvoyant*, oferind întotdeauna cele mai frecvente top k rezultate, și *Zangarle*, un model propus în [50]. Metoda noastră este cea mai eficientă pentru precizia la o recomandare (P@1), precum se poate observa în Tabelul 7.4. Zangarle totuși oferă performațe lejer mai bune pentru predicția la 3 și la 5 recomandări (P@3 and P@5).

De asemenea am făcut o analiză a cum am putea să ne îmbunătățim metodologia. Descoperim că recomandările inline au rezultate mai bune pentru valori mari ale lui k, deci când oferim mai multe recomandări, în timp ce motoda contextuală oferă rezultate mai bune pentru k mic, deci mai puține recomandări. Pentru că Zangarle și metoda contextuală sunt conceptual similare, testăm performanța lui #SHARP! înlocuind recomandările contextuale cu Similarity Rank aparținând Zangarle. Observăm, în Figurile 7.4 și 7.5, că cele două metode împreună depăsec performanțele tuturor metodelor individuale, oferind o mai bună precizie și recall.

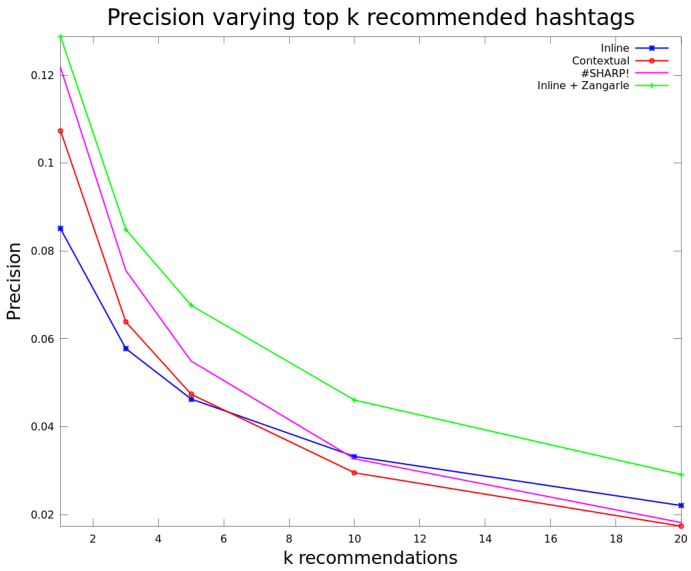

Figura 7.4: Precizia @k

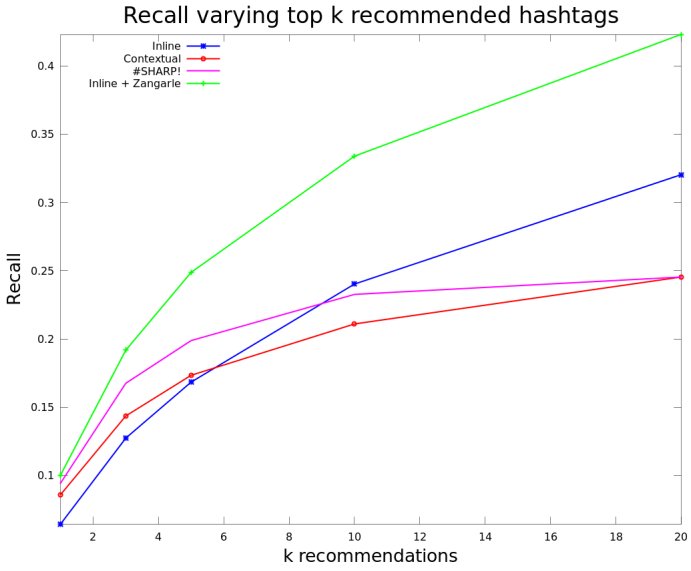

Figura 7.5: Recall-ul @k

Capitolul 8

Concluzii

8.1 Concluzii și direcții viitoare

Având în vedere obiectivele propuse anterior, ne-am concentrat atenția pe posibilele moduri de atingere a acestor obiective și pe oferirea de posibile soluții. Am obținut următoarele:

- Am oferit o definiție extinsă a Social Media si am agregat opinii din alte cercetări priving Social Media. Datorită faptului că Social Media are ca și mediu printre altele și rețele sociale, ne-am concentrat atenția pe a studia ce înseamnă o rețea socială și cum funcționează acestea. Am folosit Twitter ca studiu de caz și l-am analizat din două puncte de vedere: unul aparținând utilizatorului, iar celălalt aparținând științei analizei rețelelor sociale.

- Pentru a înțelege mecanismele platformelor de Social Media investigăm aspecte legate de experiența utilizatorilor și implicarea/stimularea acestora. Abordarea noastră a exploatat tendința ludică a utilizatorilor, și anume cum pot fi transformate produse, aplicații și servici în așa fel încât să apeleze la simțul de joacă al utilizatorilor. Gamification este o tehnică ce încearcă aplicarea de mecanici de joc unor contexte non-joc pentru a stimula și angaja utilizatorii. Am trecut în revistă principiile de bază, cum acestea pot fi aplicate unei aplicații exemplu și cum sunt aplicate în platforme Social Media ca Twitter. Rezultatele acestui studiu au fost diseminate în [30].

- Avansând în analiza rețelei Twitter studiem cele mai importante aplicații. Twitter este un mediu interesant deoarece prezintă numeroase particularități. Este considerat în acelasi timp o rețea socială cât și o platformă de micro-blogging. Lucrări în domeniul analizei rețelelor sociale

tratează Twitter ca o rețea socială, iar lucrări in domeniul analizei dei text scurt abordează Twitter ca o platformă de micro-blogging. Am explorat studiile importante privind Twitter, caracteristicile și aplicațiile acestuia.

- Pentru a analiza conținutul din Twitter trebuie mai întâi să-l fi colectat. Am explorat posibilele opțiuni: cea clasică, un crawler web ce extrage conținutul HTML al paginilor, metode diseminate în [29, 27, 28], sau folosing opțiunea de streaming pusă la dispoziție de Twitter. Am profitat de avantajul de a acea acces la fluxul de date destinat pentru scopuri de cercetare, de 10 ori mai mare decât cel normal, conținut livrat în format JSON. Am ales să creem o conxiune privată la acest flux de date și să descărcăm datele direct de la sursă, pe o perioadă de 3 săptămâni.

- După studierea datele obținute, am explorat metode privind cum să procesăm acestă cantitate mare de date. Am luat în calcul posibile tehnologii și am decis să procesăm datele în mod distribuit, folosind intrumente corespunzătoare. Am decis asupra Cascading, o librărie ce abstractizează MapReduce, construită peste Hadoop. Am implementat procese pentru curățirea datelor și filtrarea acestora. Asupra setului de date rezultant am calculat o serie de statistici pentru a întelege mai bine natura datelor. Parte din aceste unelte, procese și statistici sunt descrise în [31].

- Considerând datele la dispoziție am explorat modalități de modelare a datelor pentru a reuși să ne atingem obiectivele. Am studiat metodologii, structuri și algoritmi relevanți din Information Retrieval, Recomandări și Machine Learning pentru a reuși să implementăm aplicațiile. Decidem să folosim: 1) algoritmi de învățare supervizată și nesupervizată - abordarea ML, 2) crearea unui index inversat pentru interogarea datelor dorite - abordarea IR și 3) crearea unui predictor - abordarea probabilistică. Folosind toate acestea reușim să modelăm datele în moduri diferite, însă având mereu în vizor meta-problema.

- În cercetarea noastră ne-am propus să găsim un model valid de reprezentare a hashtag-urilor. Folosind această reprezentare am clusterizat hashtag-uri, folosind un număr variabil de clustere și am analizat cum acestea se comportă în funcție de mărimea clusterului. Motivația noastră a fost să găsim moduri automate de grupare a hashtag-urilor în funcție de contextul în care apar pentru a reuși să capturăm sensul semantic al acestora din corelația cu alte hashtag-uri sau cu termenii cei

mai frecvenți din documentele clusterizate. Rezultatele sunt prezentate în [31].

- De asemenea ne-am propus o aplicație de recomandare de hashtag-uri per tweet pentru utilizatorii Twitter. În primul rând am demostrat utilitatea acestora și cum funcționează în creșterea vizibilității unui tweet. Am observat și natura dală a acestora: inline și de context. Am modelat fiecare tip de hashtag în parte. Hashtag-urile contextuale se referă la sensul semantic al unui tweet, deci reușesc să îl încadreze într-o anumită categorie sau subiect. Am creat un index inversat pentru documentele de context ale hashtag-urilor și l-am folosit ca motor de căutare. L-am folosit interogându-l cu tweet-uri și, în funcție de similaritatea acelui tweet cu documentele virtuale indexate, am primit ca recomandări hashtag-uri ordonate descrescător în funcție de scorul de similaritate. Hashtag-urile inline au fost modelate printr-o abordare probabilistică. Am construit toate combinațiile de cuvinte dintr-un tweet de text și am căutat cele mai probabile hashtag-uri în funcție de tweet, ordonate după scorul de probabilitate. Am evaluat rezultatele și le-am confruntat cu altele similare din literatură. Articolul este încă în lucru, iar echipa de cercetare este prezentă în [12].

Direcții viitoare de cercetare

O aplicație interesantă în cazul clusterizării hashtag-urilor ce încercăm să o tratăm în dezvoltări ulterioare este clustering-ul ierarhic cu scopul construirii unei taxonomii de hashtag-uri, pornind de la general la specific. O altă aplicție interesantă este crearea unui instrument ce generează în mod automatic o explicație pentru înțelesul hashtag-ului respectiv, probabil folosind tehnici de sumarizare și modelare a temei (topic modeling).

Credem că rezultatele recomandării de hashtag-uri pot fi îmbunătățite. Dorim să experimentăm și cu alte modalități de reprezentare a hashtag-urilor. Dorim să aplicăm metode de învățare pentru determinarea optimumului de greutăți pentru combinația liniară dintre recomandări inline și contextuale, adică valoare optimă pentru α. Am creat de asemenea mai multe proprietăți ce pot fi folosite pentru a întări sau a scădea importanța unor hashtag-uri (entropia, probabilitatea unui hashtag contextual considerând un hashtag inline). Intenționăm să folosim aceste proprietăți și să aplicăm L2R[1]. Dorim să încercăm oferirea de recomandări prin clasificare, folosind și comparând performanțele a Naive Bayes, Complementary Naive Bayes și Support Vector Machines (tehnologia de vârf în clasificarea de text).

[1]Learning to rank, se referă învățarea ordonării optime.

Concluzionăm cercetarea prin a menționa că ne-am atins obiectivele propuse și am reușit să propunem soluții pertinente la acestea. Considerăm că ne-am extins mult cunoștiințele în domeniul Social Media, clusterizării de text scurt și oferirea de recomandări, totodată contribuind cu soluții inovatoare. Cu toate acestea, în acest domeniu de cercetare există multe de descoperit.

8.2 Diseminarea rezultatelor din teză

Rezultatele prezentate în această teză au fost diseminate în următoarele articole:

Articole publicate/acceptate în volume de conferințe internaționale din străinătate:

- C. I. Muntean, G. A. Morar, and D. Moldovan, *Exploring the meaning behind Twitter hashtags through clustering*, 15th International Conference on Business Information Systems, Vilnius, Lithuania, In Business Information Systems Workshops, LNBIP, vol. 127, pages 231 - 242. Springer-Verlag Berlin, 2012. (accepted for publishing)
- C. I. Muntean. *Raising engagement in e-learning through gamification.* In Proceedings of the 6th ICVL Conference, pages 323-329. Editura Universitatii Bucuresti, 2011.
- C. I. Muntean, D. Moldovan, and O. Veres. *A data mining method for accurate employment search on the web.* In Proceedings of the 2010 international conference on COMATIA, pages 123-128, World Scientific and Engineering Academy and Society, 2010.

Articole publicate/acceptate în reviste naționale:

- G.A. Morar, C.I. Muntean, and G.C. Silaghi. *Implementing and running a workflow application on cloud resources.* Informatica Economica, 15(3):15-27, 2011.

Articole publicate/acceptate în reviste internaționale:

- C. I. Muntean, D. Moldovan, and O. Veres. *A personalized classification of employment offers using data mining methods.* In International Journal of Mathematical Models and Methods in Applied Sciences, 5(4):525-532, 2011.

8.3 Lista completă de articole

Lista completă a articolelor dezvoltate pe parcursul studiilor doctorale:

Articole publicate/acceptate în volume de conferințe internaționale din străinătate:

- C. I. Muntean, G. A. Morar, and D. Moldovan, *Exploring the meaning behind Twitter hashtags through clustering*, 15th International Conference on Business Information Systems, Vilnius, Lithuania, In Business Information Systems Workshops, LNBIP, vol. 127, pages 231 - 242. Springer-Verlag Berlin, 2012. (accepted for publishing)

- C. I. Muntean. *Raising engagement in e-learning through gamification.* In Proceedings of the 6th ICVL Conference, pages 323-329. Editura Universitatii Bucuresti, 2011.

- R. Baraglia, C. Frattari, C. I. Muntean, F. M. Nardini, and F. Silvestri. *RecTour: A recommender system for tourists.* In Proceedings of the 2012 Web Intelligence Workshops, WIIAT'12, Macau, China, 2012. IEEE Computer Society. (accepted for publishing)

- R. Baraglia, C. Frattari, C. I. Muntean, F. M. Nardini, and F. Silvestri. *A trajectory-based recommender system for tourism.* In Proceedings of 2012 International Conference on Active Media Technology, AMT'12 LNCS, Macau, China, 2012. Springer. (accepted for publishing)

- C. I. Muntean, D. Moldovan, and O. Veres. *A data mining method for accurate employment search on the web.* In Proceedings of the 2010 international conference on COMATIA, pages 123-128, World Scientific and Engineering Academy and Society, 2010.

Articole publicate/acceptate în reviste naționale:

- G.A. Morar, C.I. Muntean, and G.C. Silaghi. *Implementing and running a workflow application on cloud resources.* Informatica Economica, 15(3):15-27, 2011.

- G. Morar, C. I. Muntean, N. Tomai, *An Adaptative M-learning Architecture for Building and Delivering Content based on Learning Objects*, The Second Romanian Workshop on Mobile Business, 10-11 Septembrie 2010. In Informatica Economica, Vol.10, No 1/2010, pp. 63-73.

Articole publicate/acceptate în reviste internaționale:

- C. I. Muntean, D. Moldovan, and O. Veres. *A personalized classification of employment offers using data mining methods.* In International Journal of Mathematical Models and Methods in Applied Sciences, 5(4):525-532, 2011.

Articole înaintate pentru revizionare:

- Claudio Lucchese, Cristina Ioana Muntean, Raffaele Perego, Fabrizio Silvestri, Hossein Vahabi, Rossano Venturini, *Recommendation Systems in UCG*, book chapter in Mining of User Generated Content and Its Applications to be published by Tailor & Francis (CRC Press).

Articole în lucru:

- Diego Ceccarelli, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Fabrizio Silvestri, *#SHARP! : a System for HAshtag Recommendation.*

Bibliografie

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[2] Albert Angel, Nick Koudas, Nikos Sarkas, and Divesh Srivastava. What's on the grapevine? In Ugur Çetintemel, Stanley B. Zdonik, Donald Kossmann, and Nesime Tatbul, editors, *SIGMOD Conference*, pages 1047–1050. ACM, 2009.

[3] S. Attfield, G. Kazai, M. Lalmas, and B. Piwowarski. Towards a science of user engagement (position paper). In *WSDM Workshop on User Modeling for Web Applications*, February 2011.

[4] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[5] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using wikipedia. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788, New York, NY, USA, 2007. ACM.

[6] Ranieri Baraglia, Claudio Frattari, Cristina Muntean, Franco Maria Nardini, and Fabrizio Silvestri. Rectour: A recommender system for tourists. In *Proceedings of the 2012 Web Intelligence Workshops*, WI-IAT'12, Macau, China, 2012. IEEE Computer Society. Accepted for publishing.

[7] Ranieri Baraglia, Claudio Frattari, Cristina Muntean, Franco Maria Nardini, and Fabrizio Silvestri. A trajectory-based recommender sys-

tem for tourism. In *Proceedings of 2012 International Conference on Active Media Technology*, AMT'12 LNCS, Macau, China, 2012. Springer. Accepted for publishing.

[8] Richard Bartle. *Designing Virtual Worlds*. New Riders Games, 2003.

[9] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on Twitter. In *Proceedings of the Seventh Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, July 2010.

[10] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Structural trend analysis for online social networks. *Proc. VLDB Endow.*, 4:646–656, July 2011.

[11] P.J. Carrington, J. Scott, and S. Wasserman. *Models and methods in social network analysis*. Cambridge Univ Pr, 2005.

[12] D. Ceccarelli, Muntean C.I., Nardini F.M., Perego R., and Silvestri F. #sharp! : a system for hashtag recommendation. 2012. working paper.

[13] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1185–1194, New York, NY, USA, 2010. ACM.

[14] Cheng-Tao Chu, Sang Kyun Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Andrew Y. Ng, and Kunle Olukotun. Map-reduce for machine learning on multicore. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 281–288. MIT Press, Cambridge, MA, 2007.

[15] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[16] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.

[17] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. Gamification. using game-design elements in non-gaming

contexts. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, pages 2425–2428, New York, NY, USA, 2011. ACM.

[18] Edd Dumbill. What is big data? `http://radar.oreilly.com/2012/01/what-is-big-data.html`, 2012.

[19] David R. Flatla, Carl Gutwin, Lennart E. Nacke, Scott Bateman, and Regan L. Mandryk. Calibration games: making calibration tasks enjoyable by adding motivating game elements. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 403–412, New York, NY, USA, 2011. ACM.

[20] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178, New York, NY, USA, 2010. ACM.

[21] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60:2169–2188, November 2009.

[22] J.H. Kietzmann, K. Hermkens, I.P. McCarthy, and B.S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 2011.

[23] D. Knoke and J.H. Kuklinski. *Network analysis*, volume 28. Sage Publications, Inc, 1982.

[24] C. Li and J. Bernoff. *Groundswell: Winning in a world transformed by social technologies*. Harvard Business School Press, 2011.

[25] C. Lucchese, Muntean C.I., Perego R., Silvestri F., Vahabi H., and Venturini R. Recommendation systems in ucg. *Mining of User Generated Content and Its Applications*, 2012. book chapter submitted for review.

[26] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[27] G.A. Morar, C.I. Muntean, and G.C. Silaghi. Implementing and running a workflow application on cloud resources. *Informatica Economica*, 15(3):15–27, 2011.

[28] C.I. Muntean, D. Moldovan, and O. Veres. A data mining method for accurate employment search on the web. In *Proceedings of the 2010 international conference on Communication and management in technological innovation and academic globalization*, pages 123–128. World Scientific and Engineering Academy and Society (WSEAS), 2010.

[29] C.I. Muntean, D. Moldovan, and O. Veres. A personalized classification of employment offers using data mining methods. *Internationa Journal of Mathematical Models and Methods in Applied Sciences*, 5(4):525–532, 2011.

[30] Cristina Ioana Muntean. Raising engagement in e-learning through gamification. Number 6 in 6th ICVL 2011, pages 323 – 329. Editura Universitatii Bucuresti, 2011.

[31] Cristina Ioana Muntean, Gabriela Andreea Morar, and Darie Moldovan. Exploring the meaning behind twitter hashtags through clustering. Number 127 in Lecture Notes in Business Information Systems, pages 231 – 242. Springer-Verlag Berlin, 2012.

[32] Brendan O'Connor, Michel Krieger, and David Ahn. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In William W. Cohen, Samuel Gosling, William W. Cohen, and Samuel Gosling, editors, *ICWSM*. The AAAI Press, 2010.

[33] A. Osterwalder. The business model ontology: A proposition in a design science approach. *Academic Dissertation, Universite de Lausanne, Ecole des Hautes Etudes Commerciales*, 2, 2004.

[34] A. Osterwalder, Y. Pigneur, et al. An e-business model ontology for modeling e-business. In *15th Bled Electronic Commerce Conference*, pages 17–19. Bled, Slovenia, 2002.

[35] A. Osterwalder, Y. Pigneur, and C.L. Tucci. Clarifying business models: Origins, present, and future of the concept. *Communications of the association for Information Systems*, 16(1):1–25, 2005.

[36] John Pavlus. The game of life. *Scientific American*, 303:43–44, 2011.

[37] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. 2011.

[38] E.T. Peterson and J. Carrabis. Measuring the immeasurable: Visitor engagement. *Research and Analysis from Web Analytics Demystified, the Web Analytics Thought Leaders*, 2008.

[39] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 91–100, New York, NY, USA, 2008. ACM.

[40] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 385–388, New York, NY, USA, 2009. ACM.

[41] C.K. Prahalad and V. Ramaswamy. *The Future of Competition: Co-Creating Unique Value With Customers*. Harvard Business School Pub., 2004.

[42] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51, New York, NY, USA, 2009. ACM.

[43] Don Tapscott and Anthony D Williams. *Wikinomics: How Mass Collaboration Changes Everything*, volume 58. Portfolio, 2006.

[44] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, IMC '11, pages 243–258, New York, NY, USA, 2011. ACM.

[45] M. Tsvetovat and A. Kouznetsov. *Social Network Analysis for Startups: Finding Connections on the Social Web*. Real Time Bks. O'Reilly Media, 2011.

[46] Alex H. Wang. Dont't Follow me: Spam Detection in Twitter. In *Proceedings of the International Conference on Security and Cryptography (SECRYPT)*, July 2010.

[47] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[48] Tan Xu and Douglas W. Oard. Wikipedia-based topic clustering for microblogs. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011.

[49] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In William W. Cohen and Samuel Gosling, editors, *ICWSM*. The AAAI Press, 2010.

[50] E. Zangerle, W. Gassler, and Specht G. Recommending #-Tags in Twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web 2011*, pages 62–73. CEUR-WS, 2011.