Opinion Mining

Onderzoeksplan masterscriptie

Mei 2006

Johan Stortelder s0355593 johanstortelder@student.ru.nl

Probleemstelling

Inleiding

Opinion Mining is een methode waarmee automatisch meningen (opinies) uit teksten kunnen worden gedestilleerd. Een OM systeem doorzoekt een van te voren gespecificeerde groep teksten, hierbij moet gedacht worden aan uitingen in weblogs, forums, nieuwsgroepen etc., en haalt hieruit de verschillende meningen ten opzichte van een bepaald onderwerp naar voren. Het automatisch bijhouden van opinies van een grote groep mensen, maakt het eenvoudig om trends te ontdekken en deze te monitoren. Dit biedt grote mogelijkheden voor wetenschappers en marktonderzoekers.

Huidige Opinion Mining systemen maken gebruik van signaalwoorden, woorden waarvan aangenomen wordt dat ze een negatieve of een positieve lading hebben. Er wordt simpelweg gekeken of dergelijke woorden in de buurt van een onderwerp voorkomen in de tekst. Enkel op basis van deze gegevens wordt vervolgens bepaald of de houding van de schrijver positief dan wel negatief ten opzichte van het onderwerp is. Deze methode blijkt niet erg goed te werken.

In dit onderzoek wordt volgens een andere methode te werk gegaan. Eerst worden teksten over een bepaald onderwerp (topic) gezocht waarvan verwacht wordt dat ze een opinie bevatten, met behulp van klassificatie technieken (Topic Search). Vervolgens wordt binnen dit cluster gekeken naar de verschillende meningen ten opzichte van dat topic (Opinion Filtering). Dit gebeurt met het Information Retrieval systeem PHASAR. PHASAR maakt gebruik van dependency triples die de samenhang en de rol van verschillende woorden in een zin herkent en opslaat. Met behulp van PHASAR kan gezocht worden naar bepalingen bij onderwerpen en naar welk werkwoord er bij het onderwerp hoort. Het is de verwachting dat met deze kennis ook veel te zeggen is over welke mening bij welk onderwerp hoort.

Het onderzoek is op te delen in drie delen. Om te beginnen een literatuurstudie over het efficiënt trainen van classificatie systemen en de theorie hierover valideren. Vervolgens dient er een prototype Opinion Mining systeem gebouwd te worden. Tot slot wordt dit Opinion Mining systeem aan een experiment onderworpen.

Onderzoeksvraag

Hoofdvraag

- Zorgt het gebruik van Topic Search en vervolgens Opinion Filtering voor een Opinion Mining systeem dat beter presteert dan wanneer gebruik gemaakt wordt van tot nu toe gebruikelijke methoden?

Deelvragen

- Wat is de ergonomisch meest efficiënte manier om een topic klassificatie systeem te trainen?
- Hoe moeten de verschillende componenten van het OM systeem samen werken?
- Hoe moet een de GUI van een OM systeem eruit zien?
- Hoe goed werkt het systeem? Wat zijn de (verwachte) prestaties?
- Wat is de mening van de gebruiker over een dergelijk systeem?

Verantwoording

Het monitoren van opinies van grote groepen mensen is voor bepaalde organisaties van grote waarde. Hierbij moet gedacht worden aan bedrijven die de houding van klanten ten opzichte van hun product automatisch bij kunnen houden, of aan politici of journalisten die tendensen in de maatschappij bij willen houden. Er bestaan nog geen academische of commerciële systemen waarmee Opinion Mining echt goed mogelijk is.

Theoretisch kader

Afbakening

Het onderzoek beperkt zich tot verschillende deelgebieden van de Information Retrieval. Allereerst tot classificatie, document clustering, en het trainen van dergelijke systemen. Daarnaast speelt het op linguïstische frasen gebaseerde IR systeem PHASAR een belangrijke rol. Tot slot wordt ook naar de state of the art en de verschillende methoden van Opinion Mining gekeken.

Naast IR spelen ook andere gebieden een rol. GUI-ontwerp/bouw voor het ontwikkelen van de interface van het OM systeem en statistiek voor het verwerken van de resultaten van experimenten.

Keuzes en vooronderstellingen

- GUI wordt in overleg met de begeleiders geïmplementeerd in de programmeertaal die het best past bij de bestaande systemen en interfaces waarmee OM systeem moet gaan werken.
- Iedere week is er contact met de begeleiding. Het liefst enkele malen per week.
- Er wordt deels op de universiteit en deels thuis aan het onderzoek gewerkt.
- De componenten van het systeem, LCS en PHASAR, bestaan al en dienen door middel van een GUI en enkele interfaces samen te gaan werken in een prototype OM systeem.
- De resources om een OM systeem te testen zijn niet zomaar voor handen. Het risico bestaat dat de onderbouwing van de kwaliteit (van het functioneren) van het prototype niet gebaseerd wordt op het uitvoeren van zijn daadwerkelijke taak, Opinion Mining uit resources op het www.
- Door het ontbreken van die resources bestaat eveneens het risico dat een gebruiksexperiment niet over zoekresultaten gaat maar over de kwaliteitsaspecten van de GUI.
- Er worden in dit onderzoek uitsluitend Engelstalige documenten gebruikt voor OM.

Termen en concepten

Opinion mining

Automatisch bepalen en volgen van meningen in (delen van) digitale stukken tekst. Achetereenvolgens worden de volgende processen doorlopen:

1. Topic Filtering:
    a. Interactive topic demarcation
    b. Topic classification
2. Opinion Filtering
    a. Interactive opinion demarcation
    b. Opinion classification
3. Opinion evaluation, monitoring, statistics.

PHASAR systeem

Phrase-based Accurate Search And Retrieval system

LCS

Linguistic Classification System

Topic Filtering

Selecteren van documenten uit een bepaalde verzameling die gaan over eenzelfde, gegeven onderwerp (topic).

Opinion filtering

Na Topic Filtering wordt van documenten over hetzelfde onderwerp de verdeling van de verschillende meningen over dat onderwerp bepaald.

Methode

Domein

Het hoofddomein van het onderzoek is Information Retrieval.

Operationalisatie

Onderzoekselementen

- Automatische document klassificatie en het LCS
- Frase gebaseerd zoeken en het PHASAR systeem
- Ergonomische GUI die topic filtering en opinion filtering combineert tot een OM systeem
- Experimentele validatie

Indicatoren

- Efficiëntie training classificatie systeem
- Gedestilleerde opinies

Informatieverzameling

De informatie verzameling zal grotendeels door literatuurstudie geschieden. Voor wat betreft de onderwerpen opinion mining, PHASAR, classificatie en het trainen van dergelijke systemen zal de nodige literatuur doorgewerkt moeten worden. Dit geldt in mindere mate voor de bouw van de GUI, maar ook hiervoor zullen handboeken en literatuur geraadpleegd worden.

Daarnaast zal ook overleg met de begeleiding als een belangrijke bron van informatie gelden. Iedere week zal er enkele malen contact zijn.

Er is een reeds bestaande implementatie van de GUI waarnaar gekeken kan worden. Het LCS is beschikbaar evenals een prototype PHASAR systeem.

Op het eind zal OM systeem aan een experiment onderworpen worden.

Informatieverwerking

De bevindingen van de literatuurstudie, het onderzoek, de bouw van GUI en van het experiment zullen in en scriptie worden vastgelegd.

Tijd en faseringsschema

Week	Datum	Activiteit
Week 20	15 - 21 mei	Opstarten, administratie, definitie, onderzoekplan
Week 21	22 - 28 mei	Literatuurstudie
Week 22	29 mei - 4 jun	Literatuurstudie
Week 23	5 – 11 jun	Literatuurstudie
Week 24	12 – 18 jun	Literatuurstudie
Week 25	19 – 25 jun	Literatuurstudie + valideren theorie door een synthetisch experiment met gepreklassificeerde documenten
Week 26	26 jun – 2 jul	Literatuurstudie + valideren theorie door een synthetisch experiment met gepreklassificeerde documenten
Week 27	3 – 9 jul	GUI bouw
Week 28	10 – 16 jul	GUI bouw + schrijven bevindingen
Week 29	17 – 23 jul	GUI bouw + schrijven bevindingen
Week 30	24 – 30 jul	GUI bouw + schrijven bevindingen
Week 31	31 jul – 6 aug	Uitloop GUI bouw + bijwerken scriptie
Week 32	7 – 13 aug	Uitloop GUI bouw + bijwerken scriptie
Week 33	14 – 20 aug	Uitloop GUI bouw + bijwerken scriptie
Week 34	21 – 27 aug	Keuze van engelstalige documentenverzameling met topics en opinies voor experiment
Week 35	28 aug – 3 sep	Experiment + schrijven bevindingen
Week 36	4 – 10 sep	Afschrijven scriptie
Week 37	11 – 17 sep	Afschrijven scriptie
Week 38	18 – 24 sep	Afschrijven scriptie
Week 39	25 sep – 1 okt	Afronden + voorbereiden presentatie
Week 40	2 – 8 okt	…
Week 41	9 – 15 okt	…