Ansätze zur Trenderkennung in Texten

Verarbeitung von großen textbasierten Datenmengen zur Analyse von Weak Signals

Marcel Schöneberg

marcel.schoeneberg@haw-hamburg.de

Hochschule für Angewandte Wissenschaften Hamburg (HAW)
Fakultät für Technik und Informatik
Department Informatik

Anwendungen 2 Präsentation

24.04.2014

Zusammenfassung

Ziele:

- Trends erkennen bevor sie offensichtlich werden
- Trendthemen in Twitter entdecken
- Versuch die politische Entwicklung der Piratenpartei vorherzusagen

Probleme:

- BigData (Information vs. Rauschen)
- Lösungen sind oft Domänengebunden
- Trendsuche in Text ist weniger verständlich als in Zahlenreihen
- Begriffsvielfalt
- Oft sehr viel theoretischer Hintergrund

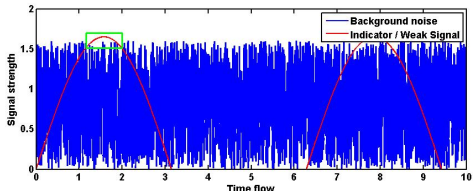

Abbildung: Illustration eines Weak Signals

1

[1] Bildquelle/siehe auch:[Sch13]

Paper 1

Deriving market intelligence from microblogs (2013)

Yung-Ming Li, Tsung-Ying Li
Institute of Information Management, National Chiao Tung University, Taiwan

[LL13]

Paper 1: Ziele

- Erkennen von aufkommenden Trends
- Klassifizierung von Tweets
- Glaubwürdigkeitseinschätzung von Äußerungen
- Zusammenfassung von Meinungen

Paper 1: Umsetzung I

- Trendy Topic Detection Module: Berechnung der Topic Tendency Score (TTS) als:

$$TTS_{q,t} = TF_{q,t} * IDF_{q,t} * MPP_{q,t}$$

  - Termfrequency: Wie häufig kommt Wort w in Dokument d vor
  - Inverse Termfrequency: Wie relevant ist ein Wort innerhalb des gesamten Korpus
    $\log \frac{|Documents|}{Term}$
  - MPP (Meronym Pattern Portion): Wie oft kommt ein Term in einer (vordefinierten) Teil-Ganzes-Beziehung vor (z.B. battery PART OF iPhone)
    $\frac{\text{Anzahl von t im Dokument}}{TF_t}$
    $\rightarrow$ Stellt eine semantische Verbindung über vordefinierte Pattern her – daher wird die Präzision der Topicerkennung erhöht.
  - Themen entsprechend des TTS sortieren.

Paper 1: Umsetzung II

→ Die n höchsten TTS zeigen die rechnerisch wichtigsten Worte eines Dokumentes an – diese können im Folgenden für Zusammenfassungen benutzt werden.

Paper 1: Bewertung

- Interessante (weiterführende) Ansätze
- Lösung über Termfrequency zunächst recht einfach – Erweiterung durch MPP ggf. nützlich
- Weitere im Paper ausgeführte Module könnten eine Trenderkennung noch interessanter machen.
- Glaubwürdigkeitsmodul benötigt zusätzliche „soziale“Daten um Zusammenhänge zwischen Personen in die Berechnung mit einfließen zu lassen (Repräsentation als Bekanntschafts-Graph)
- Klassifikation und Zusammenfassung werden später noch tiefergehend behandelt

Paper 2

Latent Dirichlet Allocation (2003)

David M. Blei
University of California (Berkeley)

Andrew Y. Ng
Stanford University

Michael I. Jordan
University of California(Berkeley)
[BNJ03]

Paper 2: Ziele

- Entwicklung eines Topic-Modells, welches (unter Anderem) Textkorpora darstellt.
- Dokumente eines Korpus einer Topic mit einer Wahrscheinlichkeit zuordnen (z.B. via Gibbs Sampling)
- Topiczusammensetzung Φ eines Dokuments bestimmen
- Keywords β für eine Topic entdecken
- $\rightarrow$ Clustering

Paper 2: Grundlagen

- Ebenfalls ein Bag-Of-Words Ansatz
- Dokument als Mischung von Topics
- Topic als Mischung von Wörtern
- Themen sind latent - d.h. von außen nicht direkt zu erkennen
- Iteratives Anpassen bis zu einem Punkt an dem nahezu keine Änderungen mehr auftreten
- $\rightarrow$ Topic-Zuordnung erlernen

Paper 2: Grundlegender Algorithmus

Algorithm 1 Grundlegendes Vorgehen beim Lernen durch Gibbs Sampling

Input: Set of topics, Corpus of documents
Output: Words assigned to topics

```
//Initiale Topicvergabe
for all <Documents in Corpus> do
    for all <Words in actual Document> do
        topicForWordInDocument ← randomTopic
    end for
end for
//Update der Themen aufgrund des Vorwissens
for all <Documents in Corpus> do
    for all <Words in Document> do
        for all <Topics> do
            //Menge von Wörtern im Dokument, die der Topic t zugewiesen sind, berechnen
            wordsAssignedToTopic ← p(topic t | document d)
            //Menge von Zuweisungen an Topic t bei aktuellem Wort (über alle Dokumente)
            //Bedenke: w kann in verschiedenen Dokumenten anderen Topics zugewiesen sein
            topicsAssignedToWord ← p(word w | topic t)
        end for
        //Neue Topic t mit Wahrscheinlichkeit max(p) zuweisen (p ≙ Wahrscheinlichkeit von t generierte w)
        newTopicForWord ← wordsAssignedToTopic ∗ topicsAssignedToWord
    end for
end for
```

Erläuterung

- Der Updateschritt nimmt an, dass alle anderen Zuweisungen von Topics korrekt sind
- Zuweisung für ein Wort basiert auf dem generativen Modell von LDA
- Nach beliebig vielen Wiederholungen des Update-Schrittes wird ein nahezu stabiler Zustand erreicht indem die Zuweisungen sich nicht mehr ändern
- $\rightarrow$ Topiczuweisungen pro Wort, Topicverteilung pro Dokument, sowie Keywords pro Topic bekannt

Paper 2: Bewertung

- Grundlage für viele aufbauende Modelle
- Theorie dieses Ansatzes komplizierter als gezeigt, allerdings reicht der Grundverständnis zunächst
- Clustering ist eine gute Grundlage zur Trenderkennung, da der Korpus komprimiert wird
- Großer Nachteil ist, dass die Clusteranzahl vorher bekannt sein muss

Paper 3

SNS-based Issue Detection and Related News Summarization Scheme (2014)

Daeyong Kim et al
School of EE
Korea University
Seoul, Korea

[KKK+14]

Paper 3: Ziele

- Entwurf einer social-network-service (SNS) Themen Erkennung
- Beziehung zwischen einzelnen Keywords aufdecken
- Zusammenfassung von mit den Funden verbundenen Nachrichten

Paper 3: Idee

Extraktion von Keywords nicht ausreichend (Mehrdeutigkeit, wenig Kontext)

1. Extrahieren von Tweets anhand von Trend-Keywords (beispielsweise mit den Piraten assoziierte Begriffe oder durch syntaktische Analyse z.B. Wörter in Anführungsstrichen)
2. Nachrichten mit relativer Begriffsnähe suchen (Begriffsüberdeckung: Gehäuftes Vorkommen von verbundenen Worten)
3. Zusammenfassung eines Themas erstellen und einen repräsentativen Tweet finden

Issue Detector I

- Extrahiere Nomen als Keywords aus einer Anzahl von Tweets (z.B. häufig vorkommend)
- Nutze nur Nomen - andere Wortarten haben eine höhere Wahrscheinlichkeit in mehreren Tweets vorzukommen
- Gruppiere Keywords nach ihrem kombinierten Vorkommen in Tweets (z.B. „Piraten“und „entern“)

→ Diese Gruppen bilden die Trends

- Idee: Häufige Wiederholungen von Wort-Gruppierungen beschreiben die selbe Thematik)
- Grenzwert für kombiniertes Vorkommen adaptiv während der Gruppierungsphase (Ergebnismenge beschränken)

→ z.B. zunächst eine Gruppe in der Wörter in 10 % der Fälle gemeinsam auftreten usw.)

Issue Detector II

- Auf Basis der verschiedenen Gruppen kann daraufhin versucht werden Ähnlichkeiten zwischen den Gruppierungen zu finden

→ Berechnung der Ähnlichkeit von Gruppen (ggf. Verschmelzung und/oder Nutzung als repräsentativer Tweet)

Paper 3: Bewertung

1. Ansatz verfolgt ebenfalls einen Ansatz der den Korpus komprimiert
2. Idee nicht nur die Häufigkeit einzelner Tweets sondern die Co-Occurence zu betrachten ist durchaus interessant

→ Verknüpft Bestandteile des Korpus miteinander und bietet die Möglichkeit nah beieinander liegende Themen zu erkennen

...towards Weak Signals Trenderkennung

- Alle Paper widmen sich der Entdeckung von aktuellen Trends
- **Ziel:** Vorzeitige Trenderkennung
- **Aktuell: Zusammenfassung** des Korpus, sowie (zusammenhängende) **Keywords**
  - → Buzzwords beschreiben aktuelle Topics
  - → **Veränderung der Schlüsselworte** über die Zeit verfolgen
  - → Zeitreihenanalyse
  - → Durch Komprimierung auf Zusammenfassung von Dokumenten (anstatt nur Keywords): **Kontextbindung** – Versuch den Kontext zu erhalten (im Gegensatz zu puren Bag-Of-Words Ideen)

Literature I

Blei, David M. ; Ng, Andrew Y. ; Jordan, Michael I.:
Latent Dirichlet Allocation.
In: **J. Mach. Learn. Res.** 3 (2003), März, 993–1022.
http://dl.acm.org/citation.cfm?id=944919.944937. –
ISSN 1532–4435

Chen, Edwin:
Introduction to Latent Dirichlet Allocation.
http://blog.echen.me/2011/08/22/
introduction-to-latent-dirichlet-allocation/

Literature II

Kim, Daeyong ; Kim, Daehoon ; Kim, Siwan ; Jo, Minho ; Hwang, Eenjun:
SNS-based Issue Detection and Related News Summarization Scheme.
In: **Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication**.
New York, NY, USA : ACM, 2014 (ICUIMC '14). –
ISBN 978–1–4503–2644–5, 114:1–114:7

Li, Yung-Ming ; Li, Tsung-Ying:
Deriving Market Intelligence from Microblogs.
In: **Decis. Support Syst.** 55 (2013), April, Nr. 1, 206–217.
http://dx.doi.org/10.1016/j.dss.2013.01.023. –
DOI 10.1016/j.dss.2013.01.023. –
ISSN 0167–9236

Literature III

Schöneberg, Marcel:
Weak Signals.
(2013).
http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2013-aw1/schoeneberg/bericht.pdf

Scheffler, Tobias ; Haider, Peter ; Prasse, Paul:
Latente Dirichlet Allokation,
http://www.cs.uni-potsdam.de/ml/teaching/ss11/st/LDA.pdf

Fragen?