Ansätze zur Trenderkennung in Texten

Anwendungen 2 Ausarbeitung

Marcel Schöneberg

marcel.schoeneberg@haw-hamburg.de

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

Zusammenfassung

Dieses Dokument stellt verschiedene Ansätze zur Trenderkennung in Texten dar. Zielsetzung ist es mit diesen Methoden große Datenmengen zu analysieren um in ihnen im Optimalfall 'Weak Signals' zu entdecken und damit aufkommende Trends vorherzusehen. Hierzu werden drei verschiedene Ansätze, inklusive der jeweiligen Bewertung aus Sicht des Autors mit Blick auf das Gesamtziel, vorgestellt. Im Anschluss wird der Zusammenhang der verschiedenen Ansätze erläutert. Abschließend werden zukünftige Ziele und Herausforderungen, mit welchen sich der Autor beschäftigen wird, vorgestellt.

1 Einführung

Das vorliegende Papier stellt verschiedene Ansätze zur Trenderkennung bzw. zum Textclustering vor. Diese sollen im Weiteren dazu genutzt werden schwache Signale [Sch13, vgl.] in einem Twitter Datenset aufzudecken und so aufkommende Trends zu entdecken und zu verfolgen. Die vorgestellten Ansätze gehen zum Teil über die reine Trenderkennung hinaus und besprechen weitere Verarbeitungen welche ebenfalls in späteren Analysen nützlich sein können. Diese weiterführenden Ideen werden ebenfalls im Rahmen ihrer Wichtigkeit vorgestellt.

2 Deriving market intelligence from mircoblogs (2013)

2.1 Ziele

Das im Folgenden vorgestellte Paper [LL13] ist konkret auf die Analyse von Mikroblogs (wie z.B. Twitter) zugeschnitten. Daher wurden laut den Autoren u.A. die Länge bzw. Kompaktheit von Nachrichten wie Tweets berücksichtigt. Ebenso wichtig war ihnen das Verteilungsmodell von Informationen (z.B. die Nutzung von Retweets in Twitter) zu bedenken.

Die Hauptziele des Papiers widmen sich der Erkennung von aufkommenden Trends welche in den Meinungen von Users ausgedrückt werden. Darüber hinaus sollen diese Meinungen klassifiziert werden, hierbei ist der Kern der Analyse zu extrahieren welchen Eindruck ein Benutzer vermitteln will. Das dritte Kernproblem welches die Autoren angehen wollen ist die 'Glaubwürdigkeitseinschätzungen"von Äußerungen. Der Zweck dieser Einschätzung soll es sein eine repräsentativere Zusammenfassung der Äußerungen liefern zu können. Diese bildet die vierte Fragestellung des Papers und soll dazu dienen einen aggregierten Eindruck aller Äußerungen zu generieren. Dieses Fazit kann daraufhin beispielsweise an Entscheidungsträger weitergereicht werden, welche u.A. darauf basierend eine Strategie zur Erreichung von z.B. Unternehmenszielen im Marketingbereich beschließen.

2.2 Umsetzung

Die folgenden Abschnitte beschreiben das Framework welches die Ziele der Autoren umsetzen soll. Dieses ist in vier Module aufgeteilt welche die jeweiligen Teilziele realisieren sollen.

2.2.1 Trendy topic detection module

Dieses Modul hat die Zielsetzung aufkommende Trends zu erkennen, hierzu wird jedem Term ein Wert zugewiesen welcher die Relevanz widerspiegeln soll. Dazu wird ein Topic Tendency Score (TTS) für einen Term t bei einer Suchanfrage q (auf einer Sammlung von

Dokumenten/ Opinons O) berechnet. Der Wert berechnet sich als

$$TTS_{\text{q,t}} = TF_{\text{q,t}} * IDF_{\text{q,t}} * MPP_{\text{q,t}} \tag{1}$$

Hierbei ist $TF_{\text{q,t}}$ die Häufigkeit mit der t im gesamten Datenset O_q vorkommt und IDF die inverse document frequency($IDF_{\text{q,t}} = log(\frac{|O|}{|O_q : t \in O_w|})$) - also ein Maß für die Bedeutung eines Termes im gesamten Dokumentenkorpus in Relation zu einzelnen Vorkommen. Der MPP-Wert in der Formel 2.2.1 stellt die Häufigkeit dar, wie oft der Suchterm innerhalb von vordefinierten Pattern auftritt. Diese Pattern definieren Meronyme (linguistische 'ist Teil von'-Relationen), welche die Präzision bei der Erkennung von Topics erhöhen sollen. Ein Beispiel für ein solches Muster könnte ein Tweet wie 'Battery of iPhone is not good' sein, hierbei würde der Ausdruck 'PART of ENTITY' greifen und man kann schlussfolgern, dass 'Battery' ein Bestandteil des iPhones ist. Die Berechnung des MPP-Wertes funktioniert über die Formel:

$$MPP_{\text{q,t}} = \frac{\text{Anzahl der Vorkommen von t in } O_q \text{ mit Pattern}}{TF_{\text{q,t}}} \tag{2}$$

Die Terme mit einem guten TTS-Wert werden als relevante Topics in die weitere Verarbeitung übernommen.

2.2.2 Opinion classification module

Dieses Modul soll die Grundrichtung einer Meinung bzw. eines Tweets, sowie dessen Subjektivität ermitteln. Diese Analyse hat zunächst nur einen begrenzten Wert für den Autor dieses Papers, da das Kernziel die Trenderkennung ist, daher wird dieses Modul nur sehr kurz angerissen.

Die Grundidee des 'opinion classification' Modules ist zum einen eine 'subjectivity analysis' durchzuführen. Diese basiert auf der Annahme, dass subjektivere Tweets eine höhere Dichte von emotionalen, sowie sentimentalen Worten aufweisen. Daher werden diese Wörter anhand eines vordefinierten Sets (inklusive entsprechender Synonyme etc.) abgezählt und in Relation zu den anderen vorkommenden Wörtern innerhalb eines Tweets gesetzt.

Der zweite verfolgte Ansatz der Autoren ist eine 'sentiment classification' durchzuführen. Ziel hierbei ist es eine Äußerung in die Cluster positiv bzw. negativ einzuordnen (also z.B. eine grundsätzlich eher positive Meinung über ein Produkt oder eine ablehnende). Die Autoren des Papers arbeiten hierbei mit einer antrainierten Support Vector Maschine, also einer Methode des überwachten Lernens.

2.2.3 Credibility assessment module

Das dritte Modul des von den Autoren vorgestellten Frameworks ist das 'credibility assessment module'. Dieses soll seriöse Quellen von weniger nützlichen unterscheiden um glaubwürdigere Informationen gegebenenfalls anders zu gewichten. Hierbei untersuchen

die Autoren zwei Faktoren: Die Glaubwürdigkeit des Autors einer Aussage, sowie den entsprechenden Inhalt. Allerdings dient auch diese Analyse nicht im Kern der Trenderkennung und wird daher vom Autor dieses Papers nur kurz umrissen.

Die Glaubwürdigkeit eines Tweet-Autors berechnen die Paperautoren durch:

$$Glaubwrdigkeit_{\text{User}} = \frac{\text{Anzahl von Followern}}{\text{Anzahl der User denen der Autor folgt}} \tag{3}$$

Dieser Gedanke basiert also auf der Annahme, dass die Anzahl der Follower (wer ist am Inhalt des Autors interessiert) bzw. die Menge der Leute an denen eine Person Interesse zeigt ausschlaggebend für die Glaubwürdigkeit eines Users ist.

Der Inhalt von veröffentlichten Tweets wird ebenfalls auf seine Glaubwürdigkeit hin analysiert. Dieses geschieht, laut den Autoren, mit Hilfe der Anzahl der Tweets, welche ein User veröffentlicht bzw. retweetet (den Inhalt eines anderen Users 'hervorheben'). Dieser Glaubwürdigkeitswert berechnet sich als:

$$Glaubwrdigkeit_{\text{Tweet}} = \frac{\text{Anzahl von Tweets die vom User in Zeitperiode x geretweetet wurden}}{\text{Anzahl der Nachrichten eines Users in Zeitraum x}} \tag{4}$$

Die beiden erwähnten Glaubwürdigkeitswerte werden anschließend verrechnet:

$$CS = \sqrt{Glaubwrdigkeit_{\text{User}} * Glaubwrdigkeit_{\text{Tweet}}} \tag{5}$$

und bilden so die Gesamtglaubwürdigkeit.

2.2.4 Numeric summarization module

Dieses Modul verrechnet die bis hierhin gewonnen Informationen um einer Äußerung einen Wert zuzuweisen, welcher die Glaubwürdigkeit und die Objektivität widerspiegelt. Dieser Wert ist laut den Autoren nützlich um Meinungen und deren Entwicklungen in sozialen Medien zu verfolgen, da er es erlaubt subjektive, sowie unglaubwürdige Posts von anderen zu separieren und so eine Grundlage für weitere Analysen zu schaffen.

2.3 Bewertung

Betrachtet man die vorgestellten Ansätze des Papers so werden interessante Ideen für die Verarbeitung von Nachrichten in social Media geliefert. Diese können im Prinzip nützliche Informationen liefern, allerdings ist der Autor dieses Papers der Meinung, dass die Einzelmodule zum Teil des rudimentär beschrieben sind. Insgesamt wird das Kernthema des Autors - die Trenderkennung - nicht stark genug hervorgehoben, dennoch werden Grundprinzipien wie Bewertungsfunktionen (TF, IDF) sehr anschaulich genutzt und um linguistische Aspekte erweitert. Dieses zeigt, dass sprachliche Faktoren in weiteren Analysen zur Trenderkennung nicht vernachlässigt werden sollten. Ebenso sind die Ideen der anderen Module für spätere Analysen grundsätzlich interessant, da sie das Grundrauschen

welches aufkommende Trends verdeckt (z.B. Spam) bekämpfen können. Abschließend ist das Paper daher eine Ideensammlung und zeigt einige Aspekte auf welche zu bedenken sind.

3 Latent Dirichlet Allocation (2003)

Im folgenden Abschnitt wird ein Clusteringverfahren vorgestellt, welches auf der latenten Dirichlet Allocation [BNJ03] basiert. Dieses ist ein Basisverfahren welches im Laufe der Zeit um verschiedene Aspekte erweitert wurde und in verschiedenen Bibliotheken implementiert ist (siehe z.B. http://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf).

3.1 Einführung und Ziele

Die latente Dirichlet Allocation ist ein generatives Wahrscheinlichkeitsmodell für einen Korpus (z.B. eine Sammlung von Textdokumenten). Die grundlegende Idee ist, dass jedes Dokument aus einer Anzahl von Topics (auch Themen genannt) besteht (welche nach außen nicht direkt sichtbar (d.h. latent) sind). Weiterhin ist jede Topic als eine Mischung von Wörtern anzusehen, diese bilden das Thema.

In diesem Modell sollen daher die verschiedenen Wörter (und letztendlich auch die entsprechenden Dokumente) eines Korpus mit möglichst hoher Wahrscheinlichkeit einem Thema zugeordnet werden. Die nun zugeordneten Themen bilden die späteren Cluster. Basierend auf der Zuordnung von Wörtern und Dokumenten zu Themen (und damit Clustern) lässt sich die Topiczusammensetzung eines Dokumentes bestimmen (beispielsweise: 20% Topic A, 70% Topic B sowie 10% Topic C). Darüber hinaus kann man anhand der Keywords pro Cluster (z.B. die am häufigsten benutzten Worte) Schlagwörter ermitteln welche approximiert den Inhalt eines Topics (bzw. Clusters) wiedergegeben.

Grundlegend basiert die Themenzuordnung von LDA auf einem Lernverfahren welches auf baiyescher Statistik fußt und den Methoden des unüberwachten Lernens zuzuordnen ist. Weiterhin ist die Grundidee ein Bag-Of-Words Ansatz, welcher ein Dokument nur als Ansammlung von Wörtern, allerdings ohne Semantik, ansieht.

3.2 Umsetzung

Nachdem die Basisidee von LDA erklärt wurde widmet sich der nächste Abschnitt der grundlegenden Idee des Lernverfahrens welches die Topiczuordnung übernimmt.

Das Erlernen der Topicverteilung über einen Korpus ist ein Problem der bayischen Statistik. Eine Möglichkeit dieses anzugehen ist der Algorithmus des "Gibbs-Sampling"[Gri, vgl.]. Die Idee hinter dem Gibbs-Algorithmus ist es sich einer wahrscheinlichen Topicverteilung iterativ anzunähern. Hierzu wird einer ausgewählten Variable (z.B. einem Wort) ein Wert zugewiesen, dieser basiert auf ihrer bedingten Wahrscheinlichkeit in Bezug auf die anderen (bekannten) Wörter. Dieses Vorgehen basiert auf Markov-Ketten Monte Carlo

(MCMC) Methoden, welche Stichproben aus Wahrscheinlichkeitsverteilungen ermitteln. Das Ergebnis der hierbei entstehenden Markov-Kette konvergiert nach beliebig vielen Schritten mit der gesuchten (Topic)verteilung.

Abbildung 3.2 zeigt die Topicermittlung mittels Gibbs-Sampling im LDA Algorithmus in Pseudocode.

Algorithm 1 Grundlegendes Vorgehen beim Lernen durch Gibbs Sampling

```
Input: Number 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
        //Verhältnis von Wörtern des Dokumentes die Topic t zugewiesen sind
        ProportionOfWordsAssignedToTopic ← p(topic t | document d)
        //Verhältnis von Zuweisungen des aktuellen Wortes an Topic t (über alle Dokumente)
        //Bedenke: w kann in verschiedenen Dokumenten anderen Topics zugewiesen sein
        ProportionOfAssignmentToTopicOverallFromWord ← p(word w | topic t)
      end for
      //Neue Topic t mit Wahrscheinlichkeit p zuweisen (p ≙ Wahrscheinlichkeit von t generierte w)
      newTopicForWord ← wordsAssignedToTopic * topicsAssignedToWord
    end for
  end for
```

Wie im Pseudocode zu sehen ist erhält der Algorithmus als Eingabe einen Korpus, sowie die Anzahl von gewünschten Cluster/Themen. Dieses ist allerdings nur eine Zahl, sodass die gefunden Cluster später keine inhärente Semantik aufgrund des Algorithmus besitzen. Nach der initialen zufälligen Topiczuordnung für jedes Word in jedem Dokument des Korpus folgt der Updateschritt. Dieser berechnet die beiden Wahrscheinlichkeiten

$$p(\text{topic t} \mid \text{document d}) = \frac{\text{count(words assigned to topic t, document)} + \alpha}{\text{count(topics assigned to document d)} + K * \alpha} \tag{6}$$

sowie

$$p(\text{word w} \mid \text{topic t}) = \frac{\text{count(instances of word w assigned to topic t)} + \beta}{\text{count(words assigned to topic t)} + W * \beta} \tag{7}$$

hierbei sind α und β Hyperparameter, welche dem Feintuning des Algorithmus dienen. Mit den berechneten Wahrscheinlichkeiten wird dem aktuellen Wort eine neue Topicwahrscheinlichkeit zugewiesen. Die entstehenden Cluster basieren letztendlich auf der Idee, dass gemeinsam in einem Dokument vorkommende Wörter eine höhere Wahrscheinlichkeit haben einer gemeinsamen Topic anzugehören. Der für diese Erklärung benötigte mathematische Hintergrund übersteigt allerdings sowohl den Rahmen dieses Papers als auch das bis jetzt gewonnene Verständnis des Autors.

Nach Abschluss des Algorithmus können die jeweiligen Topiczuweisungen, die Topicverteilung sowie ggf. Keywords pro Topic abgelesen werden.

3.3 Fazit

Das Modell der latenten Dirichlet Allocation bildet die Grundlage für eine Reihe von auf ihm aufbauenden Modellen. Diese umfassen z.B. Author-Topic Modelle [RZGSS04], sowie Erweiterungen um Userpofile [GK04]. Auch aus diesem Grund ist es von Vorteil zunächst die Grundlagen verstanden zu haben. Darüber hinaus ist zu bedenken, dass die Theorie dieses Ansatzes durchaus komplizierter ist als dargestellt. Von einer Einführung in die zu Grunde liegende Dirichlet-Verteilung wurde abgesehen, ebenso wurde nicht näher auf die im Algorithmus verwendeten Hyperparameter eingegangen.

Betrachtet man die Vor- und Nachteile des Algorithmus so fällt zunächst auf, dass als Eingabeparameter u.A. die Anzahl der Topics benötigt wird. Dieses Vorwissen ist allerdings bei einem völlig unbekannten Korpus in aller Regel nicht vorhanden. Daher kann die Anzahl der zu erzeugenden Topics nur willkürlich festgelegt werden, bzw. bei mehreren Durchläufen verglichen werden. Dieses ist nach Ansicht des Autors zunächst ein Hindernis, welches überwunden werden muss. Grundlegend ist darüber hinaus anzumerken, dass die mathematische Basis des Verfahrens recht komplex ist. Die Anwendung des Algorithmus ist dank diverser vorhandener Implementierungen vergleichsweise einfach, allerdings ist es schwer (unerwartet schlechte) Ergebnisse zu erklären ohne den komplizierten statistischen Hintergrund komplett zu verstehen.

Insgesamt bietet die Grundidee des LDA Verfahrens dennoch große Vorteile, da ein großer Korpus im Optimalfall in kleinere Einheiten aufgeteilt wird (die Cluster/Topics). Diese sind im besten Fall leichter zu verarbeiten, da sie einen inhaltlichen Zusammenhang haben, sowie Schlagwörter (z.B. durch Häufigkeit festgelegt) welche ihren Inhalt repräsentieren. Diese Resultate können eine wichtige Grundlage zur Weiterverarbeitung und Trenderkennung sein.

Praktische Ergebnisse

Der Autor dieses Papers hat ebenso praktische Versuche auf einem Twitterdatenset mit ca 27100 Datensätzen (zum Thema Piratenpartei) durchgeführt. Die Ergebnisse dieser Arbeit sind allerdings momentan nur beim Autor verfügbar, sollen an dieser Stelle allerdings trotzdem in alle Kürze wiedergegeben werden.

Nach einer Vorverarbeitung (Datenbereinigung, verschiedene Gewichtungen, sowie Untersuchung eines auf Nomen reduzierten Korpus) wurde das LDA/Gibbs-Verfahren auf einen 'normalen', sowie auf einem nomenbasierten Korpus angewendet. Untersucht wurde die Topicverteilung (prozentualer Anteil der einzelnen Cluster am Korpus) über den jeweiligen Gesamtkorpus in Abhängigkeit von verschiedenen Gewichtungsmaßen (TF und TF-IDF), sowie die semantische Aussagekraft der Ergebniscluster/Topics. Grundsätzlich zeigten sich vergleichsweise negative Resultate. Zunächst zeigte sich dass die verschiedenen Gewichtungsmaße wenig erfolgversprechend sind und daher wenig dazu beitragen die Analysen positiv zu beeinflussen. Ein auf Nomen basierter Korpus liefert im Vergleich zu einem 'normalen' Korpus zwar durchaus bessere Stichworte innerhalb

der gefundenen Cluster allerdings fehlt diesen der Gesamtzusammenhang, so dass eine semantische Zuordnung der Cluster ebenfalls schwer fällt. Betrachtet man die Topicverteilung auf den beiden Korpora, so stellt sich diese sowohl im Nomenkorpus als auch im Normalkorpus nahezu gleich dar. Auffällig ist jeweilig die starke Abgrenzung der am stärksten frequentierten Themen zu den nächst kleineren.

Insgesamt ist die durchgeführte Untersuchung nicht so aussagekräftig wie erhofft, da eine semantische Zuordnung der einzelnen Themencluster aufgrund ihrer Schlagworte kaum gelingen kann. Eine genauere Analyse des Versuchs ist unter [Sch14] zu finden.

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

4.1 Ziele

Dieses Paper [KKK+14] hat das Ziel eine Themenerkennung für 'social-network-services' (SNS, am Beispiel von Twitter) zu entwickeln und Beziehungen zwischen gefundenen Keywords aufzudecken. Weiterhin sollen die Nachrichten welche mit den Funden verbunden sind für Endbenutzer zusammengefasst werden um eine kompakte Übersicht zu generieren.

Hierzu werden von den Autoren zunächst anhand der Top Twitter Keywords (Hashtags) Tweets gesammelt, als nächsten werden aus dieser Sammlung Trendthemen ermittelt. Für diese verschiedenen Themengebiete wird jeweils ein repräsentativer Tweet gesucht und entsprechend der Trends offizielle Nachrichten aus anderen Quellen (z.B. Google News) gefunden und zusammengefasst.

4.2 Umsetzung

Das von den Autoren entwickelten System besteht aus vier Komponenten.

Die erste ist der **'trend analyzer'** welcher entsprechende Keywords, sowie semantisch ähnliche Begriffe aus der Datenbasis extrahiert. Die nomenbasierte Keyworderkennung erfolgt hierbei nach einfachen syntaktischen Regeln (Großschreibung bzw. in Anführungsstriche gesetzte Wörter). Darüber hinaus werden verschiedene, nicht näher erläuterte, Heuristiken benutzt um Synonyme, Tippfehler usw. zu erkennen und damit Duplikate zu eliminieren. Weiterhin werden verschiedene Keywords in Verbindung gesetzt sofern sie im selben Tweet vorkommen - basierend auf der Überlegungen, dass in diesem Fall auch eine semantische Verbindung zwischen ihnen besteht.

Der **'issue detector'** sammelt mit den Keywords verbundene Tweets und extrahiert Trends anhand des gemeinsamen Vorkommens von Nomen. Dieser Schritt der Verarbeitung gruppiert die gesammelten Nomen anhand der Rate ihres gemeinsamen Vorkommens innerhalb eines Tweets. Die Gruppen bilden ein Themengebiet und basieren auf der auf dem häufigeren Vorkommen der Keywords, sowie einem berechneten Wert welcher die

Zusammengehörigkeit von Tweets darstellt. Dieser berechnet sich u.a aus der Anzahl von Tweets welche möglichst viele der ermittelten Keywords beinhalten.

Als nächster Verarbeitungsschritt sucht der **'issue summarizer'** möglichen Wortkandidaten für eine Trendzusammenfassung und wählt anschließend einen repräsentativen Tweet. Hierbei werden auch Verben als mögliche Zusammenfassungskandidaten in Betracht gezogen, da diese Kontextwissen vermitteln. Das Zusammenstellen der Liste von Wortkandidaten, welche die Autoren nutzen wird nicht detailliert beschrieben. Zusätzlich zu der Liste wird pro Thema ein repräsentativer Tweet gewählt. Dieser wird anhand verschiedener Parameter berechnet. In die Berechnung der Autoren fließen hierbei unter anderem Worthäufigkeiten, Anzahl von Retweets, sowie die Menge von zusammen vorkommenden Nomen mit ein. Der so berechnete Tweet soll dazu dienen einem späteren Lesen das Verständnis eines gefunden Themengebietes zu erleichtern.

Zuletzt nutzt der **'news summarizer'** die Ergebnisse um Nachrichten (aus Zeitungen etc.) zu den Keywords zu finden. Aus diesen Nachrichten werden Kernsätze extrahiert und als Zusammenfassung genutzt. Hierbei werden von den Autoren aktuellere Nachrichten bevorzugt. Die Kerninhalte werden hierbei wie folgt extrahiert: Zunächst wird der Nachrichtenartikel in Abschnitte gegliedert, daraufhin wird für jeden Abschnitt ein Wert berechnet welcher vor allem auf dem Vorkommen der Keywords basiert. Der Abschnitt mit dem besten Wert wird als Zusammenfassung des gesamten Artikels gewählt.

4.3 Bewertung

Insgesamt bietet das Paper mit seinen Modulen interessante Ansätze zur Trenderkennung, sowie deren Zusammenfassung. Die Wahl eines repräsentativen Tweets, sowie die Nutzung von externen Nachrichtenquellen zur Anreicherung und zur Erhöhung des Verständnisses eines gebildeten Clusters sind ggf. nützliche Ideen. Dieses zeigt sich vor allem im Vergleich mit anderen Clusteringverfahren (z.B. LDA mit Gibbs-Sampling), welche für eine semantische Einordnung nur eine Liste mit Keywords zur Verfügung stellen. Allerdings sieht der Autor dieses Papers Probleme bei der Erstellung der Keywordlisten, welche für die Nachrichtenzusammenfassung genutzt werden. Sind diese nicht aussagekräftig, so ist auch die Anreicherung aus Print- bzw. Onlinenachrichten gefährdet. Darüber hinaus sind die entscheidenden Kernbereiche des beschriebenen Papers oft sehr vage beschrieben. Zwar existieren diverse Formeln, deren Grundgedanke wird allerdings oft nur sehr gering beleuchtet. Dieses mindert den Gesamteindruck des Papers.

5 Zusammenhang

Die in dieser Ausarbeitung vorgestellten Paper stellen teils verschiedene Themengebiete in den Fokus, sie alle haben allerdings einen gemeinsamen Nenner: Die Erkennung von Trends bzw. Clustering von Inhalt zu Gruppen welche Themen darstellen. Das ursprüngliche Ziel des Autors war es allerdings 'Weak Signals' / schwache Signale zu untersuchen [Sch13, vgl.]. Mit ihrer Hilfe sollten **aufkommende** Trends erkannt werden.

Dieses Ziel gilt auch weiterhin. Die vorgestellten Verfahren bilden z.T. eine gute Grundlage, bzw. liefern Ideen wie dieses Ziel zu erreichen ist. Die vorgestellten Ansätze reduzieren einen gegebenen Korpus auf verarbeitbare Featuresets (Cluster sowie Keywords) und arbeiten so Trends heraus. Der Autor ist der Meinung, dass auf diese Weise auch eine gute Grundlage geschaffen wird um aufkommende Trends zu erkennen. Diese sind allerdings schwächer ausgeprägt, haben aber ähnliche Eigenschaften (wie z.B eine Häufung von Schlagwörtern welche einen gemeinsamen Kontext haben). Ziel muss es also sein kleine Gruppen/Cluster welche ein gemeinsames Thema haben zu erkennen und ihre Veränderung über die Zeit (das Entstehen eines Trends) zu verfolgen und zu erkennen. Die Grundlagen hierfür sind nun gelegt.

6 Zusammenfassung und Ausblick

Dieses Paper hat eine Übersicht über verschiedene Ansätze der Trenderkennung, des Textclustering, sowie weiterführender Methoden zur Verarbeitung der Ergebnisse geliefert. Hierbei zeigte sich, dass eine Erkennung von Trends durch Gruppierung von ähnlichen Inhalten möglich ist. Ein weiterer Kernpunkt der hervorsticht ist, dass vor allem das Zuordnen von Semantik zu einem Cluster ein großes Problem darstellt. Dieses muss allerdings gelöst werden, da ein Cluster ohne eine für Menschen ersichtliche Bedeutung wertlos ist.

Darüber hinaus zeigen die vorgestellten Ansätze, dass weiterführende Überlegungen wie Glaubwürdigkeitseinschätzungen oder die Anreicherungen der Ergebnisse aus nicht-primären Quellen (Zeitungen, selbst generierte Zusammenfassungen) sehr wichtig sind. Diese können zum einen die Rohdaten verfeinern, sowie die Verständlichkeit der Ergebnisse massiv verbessern.

Grundsätzlich zeigt sich, dass am Thema der Trenderkennung durchaus geforscht wird. Allerdings sind die Ergebnisse und Umsetzungen aus Sicht des Autors oft ungenügend beschrieben. So wird z.T. die grundlegende Idee nur grob umrissen, so dass es kompliziert ist den Kern des Lösungsansatzes zu verstehen und zu bewerten (dieses auch weil die Lösungen oft auf bestimmte Szenarien wie soziale Netzwerke oder enge Themengebiete zugeschnitten sind).

Diese Arbeit widmete sich vor allem der Erkennung von aktuellen Trends und Themen. Der nächste Schritt muss die Verfolgung von Themen über die Zeit sein. Daher wird der Autor sich als nächstes dem Gebiet der Zeitreihenanalyse widmen und dieses mit den vorgestellten Ansätzen (und Erfahrung in ihrer Anwendung und Stärken sowie Schwächen) verbinden. Die erfolgreiche Erfüllung dieses Ziels wird allerdings durch die mäßigen praktischen Ergebnisse (u.A. Nutzung des LDA/Gibbs Clusterings) auf einem Testdatenset erschwert. Aus diesem Grund muss das Ziel bzw. dessen Lösungsstrategie ggf. überdacht werden.

Literatur

[BNJ03] 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

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

[GK04] Girolami, Mark ; Kabán, Ata: Simplicial mixtures of Markov chains: Distributed modelling of dynamic user profiles. In: **In Advances in Neural Information Procesing Systems 16**, MIT Press, 2004, S. 9–16

[Gri] Griffiths, Tom: Gibbs sampling in the generative model of Latent Dirichlet Allocation. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.138.3760&rep=rep1&type=pdf

[KKK+14] 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

[LL13] 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

[RZGSS04] Rosen-Zvi, Michal ; Griffiths, Thomas ; Steyvers, Mark ; Smyth, Padhraic: The Author-topic Model for Authors and Documents. In: **Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence**. Arlington, Virginia, United States : AUAI Press, 2004 (UAI '04). – ISBN 0–9749039–0–6, 487–494

[Sch13] Schöneberg, Marcel: Weak Signals. (2013). http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2013-aw1/schoeneberg/bericht.pdf

[Sch14] Schöneberg, Marcel: Textclustering durch Methoden des maschinellen Lernens. (2014)

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