Topic Time

Matthias Querbach

Universität Koblenz-Landau
Campus Koblenz
Institute for Web Science and Technologies

1 Einleitung

„Erkläre die Vergangenheit, erkenne die Gegenwart, sage die Zukunft voraus." Bereits Hippokrates (460 - 377 v. Chr.) erstrebte dies über 2000 Jahre vor unserer Zeit.
Die Gegenwart zu erkennen scheint einfach zu sein. Sie ist ist allgegenwärtig, wird für uns tagtäglich in Zeitungsartikeln oder den Nachrichten aufbereitet. Die Vergangenheit zu rekapitulieren ist durchaus möglich, bedarf aber eines großen Aufwands, da alle Informationen, die heute die Gegenwart widerspiegeln, bereits morgen Vergangenheit sind. Die Menge an Informationen ist unendlich. Noch schwieriger ist es die Zukunft präzise vorherzusagen.
Ching-man Au Yeung und Adam Jatowt machten es sich jedoch zur Aufgabe genau dies zu tun, eine Aufbereitung der Vergangenheit und eine Vorhersage der Zukunft. Ihr Ziel ist es aus sehr großen Textsammlungen, wie z.B. eine Sammlung von Zeitungsartikeln, Informationen zu Zeitpunkten und die dann stattfindenden Ereignisse zu gewinnen. Hierbei handelt es sich um Schlagwörter, die das Ereignis beschreiben, und Datumsangaben.
Im Folgenden wird die Vorgehensweise von Yeung und Jatowt in Kapitel 3 beschrieben. Sie verwenden dazu Verfahren und Gleichungen, die in den Kapiteln 4 und 5 erläutert werden. Anschließend erfolgt eine kurze Zusammenfassung ihrer Ergebnisse in Kapitel 6.

2 Notation und Terminologie

In dieser Ausarbeitung werden immer wieder Begriffe aus dem Sprachgebrauch von Textsammlungen, wie z.B. „Wort", „Dokument", „Korpus" und „Topic" erwähnt. In Anlehnung an [2, S. 995] werden die Begriffe wie folgt definiert:

- Ein *Wort* w ist die Basiseinheit der diskreten Daten.
- Ein *Dokument* d ist eine Sequenz aus N Wörtern, angegeben durch $W = (w_1, ..., w_i, ..., w_N)$, wobei w_i das $i$te Wort der Sequenz ist.
- Ein *Korpus* c ist eine Sammlung von M Dokumenten, angegeben durch $c = (d_1, ..., d_i, ..., d_M)$.
- Ein *Topic* z beschreibt ein Themengebiet, Thema, Ereignis o.Ä.. Einem Korpus c sind K verschiedene Topics zugeordnet.

Ein Wort w ist mit einer bestimmten Wahrscheinlichkeit einem Topic z zugeordnet. Ein Wort kann beliebig vielen Topics zugeordnet werden.
Ein Topic z ist mit einer bestimmten Wahrscheinlichkeit einem Dokument d zugeordnet. Ein Topic kann beliebig vielen Dokumenten zugeordnet werden.
Es handelt sich hierbei um Multinominalverteilungen.

3 Vorgehensweise

Ching-man Au Yeung und Adam Jatowt untersuchen große Textkorpora, die temporale Ausdrücke[1] enthalten. Ziel ihrer Arbeit ist es diese Dokumente zum einen auf

a) Verweise in die Vergangenheit [1] sowie auf
b) Informationen, die sich auf die Zukunft beziehen [5],

zu untersuchen. Yeung und Jatowt bedienen sich in beiden Fällen einer ähnlichen Vorgehensweise, die nachfolgend beschrieben wird. Es folgt eine Zusammenfassung der oben genannten Artikel.
Zunächst erfolgt die Sammlung der Testdaten, beschrieben in Kapitel 3.1. Anschließend erfolgt ein Pre-Processing, in welchem die Daten von nicht verwendeten Informationen befreit werden, beschrieben in Kapitel 3.2, sowie die Extraktion temporaler Ausdrücke, d.h. die Suche nach Verweisen auf Zeitpunkte in die Vergangenheit oder Zukunft, Kapitel 3.3. Anschließend kann die Textanalyse, die Untersuchung der Verbindungen innerhalb der Textsammlung, näher erläutert in Kapitel 3.4, vorgenommen werden.

3.1 Sammlung der Daten

Für die Auswertung wurde eine große Menge an Daten zusammengetragen. Hierfür wurden alle Artikel von Google News Archive[2] einbezogen, die den Suchkriterien entsprachen. Die Suchanfragen waren im Fall a) 32 Ländernamen, da diese häufig mit anderen Topics in Verbindung gebracht werden können. Im Fall b) wurden 61 unterschiedliche Anfragen aus den Kategorien Ländernamen, Firmennamen, Personen und „anderes“ durchgeführt. Es wurden in beiden Fällen nur Artikel, die in der Zeitspanne von 1990 - 2010 veröffentlicht wurden, einbezogen, da Artikel mit einem früheren Veröffentlichungsdatum häufig nur als Bilddatei vorliegen und somit nicht betrachtet werden können.
Für Artikel, die komplett verfügbar waren, wurde von Yeung und Jatowt die komplette Webseite heruntergeladen. Bei Artikeln, die nicht komplett verfügbar waren, wurden die Kurzfassungen gespeichert. Alle Daten wurden in einer Datenbank gespeichert. Der Datensatz für Verweise in die Vergangenheit beinhaltet 2,4 Millionen, für Verweise in die Zukunft 3,6 Millionen Artikel.

[1] Mit temporalen Ausdrücken sind Ausdrücke gemeint, die Zeitangaben beinhalten.
[2] `http://news.google.com/archivesearch`

3.2 Pre-Processing

Da die Artikel als Webseite gespeichert wurden, wurden nun HTML-Tags, JavaScript Code und andere nicht inhaltliche Elemente entfernt.
Es wurde außerdem der größte Textabschnitt des Textes extrahiert. Dies ermöglichte es sich auf die relevanten Inhalte zu konzentrieren und Zeitstempel, wie z.B. in Kopf- und Fußzeilen, zu ignorieren.
Weiterhin wurden in beiden Fällen nur englischsprachige Artikel einbezogen. Anderssprachige Artikel wurden mit Hilfe eines auf n-Gramm Matching basierten Algorithmus [3] aussortiert.

3.3 Extraktion temporaler Ausdrücke

Ein weiterer wichtiger Teil des Pre-Processings ist die Extraktion temporaler Ausdrücke. Es muss erfasst werden, ob ein Dokument einen Zeitpunkt in der Vergangenheit oder in der Zukunft erwähnt und mit welchem Topic dieser Zeitpunkt verknüpft ist.
Hierfür wurde der GUTime Tagger [6] verwendet. Die Daten wurden mit der TimeML Markup Language gespeichert. GUTime Tagger kann sowohl absolute als auch relative temporale Ausdrücke identifizieren. Absolute Ausdrücke sind eindeutig einem Zeitpunkt oder -intervall zugeordnert (z.B. „09. Mai 2012“, „Oktober 1986“). Relative Ausdrücke (z.B. „vor 10 Jahren“, „in 5 Monaten“) benötigen einen Zeitpunkt als Referenz, einen sogenannten Anker, um in einen absoluten Zeitpunkt umgewandelt zu werden. Bei den betrachteten Artikeln ist dieser Anker das Erscheinungsdatum des Artikels.
Für Verweise in die Vergangenheit wurde als Granularität der Zeit ein Jahr als Einheit gewählt. Somit werden Jahres-, Monats- und auch Tagesangaben einem einzigen Wert zugeordnet, falls die Jahreszahl identisch ist. Es wurde weiterhin lediglich die Zeitspanne von 1900 - 1989 betrachtet.
Für Verweise in die Zukunft wurden als Granularität der Zeit Jahre, Monate und Tage als Einheit gewählt, abhängig von der kleinsten Einheit der Referenz. Außerdem wurden nicht nur Zeitpunkte sondern auch Zeitintervalle in die Untersuchung mit aufgenommen. Somit konnten Monats- oder Jahresangaben sowie Ausdrücke wie „nach“, „zwischen“, „innerhalb“, „ab“ und „um“ repräsentiert werden. Es wurden ferner die Jahreszeiten festen Zeitintervallen zugeordnet und Ausdrücke wie „zu Beginn“, „Mitte“ oder „am Ende“ modelliert.

3.4 Textanalyse

Um die Topics innerhalb des Textkorpus herauszufiltern wurde für Verweise in die Vergangenheit Latent Dirichlet Allocation verwendet. Es wird in Kapitel 4.1 näher erläutert. Anschließend wurden die gefundenen Topics mit der zeitlichen Verteilung verknüpft. Das Vorgehen wird in Kapitel 4.2 beschrieben. Für Verweise in die Zukunft wurde ein zu LDA ähnliches Modell verwendet, das sowohl über die Topics als auch über die Zeitverteilungen gruppiert. Dieses Verfahren wird in Kapitel 5 beschrieben.

LDA erstellt eine Wahrscheinlichkeitsverteilung von Wörtern und Topics, basierend auf einem Textkorpus. Hierfür wurde die in Kapitel 3.1 beschriebene Datensammlung weiter verfeinert. Zu jedem temporalen Ausdruck wurde jeweils nur der Satz, in dem er vorkommt, sowie der vorherige und der darauffolgende Satz in einem neuen Dokument gespeichert. Dies ist der Kontext des Ausdrucks. LDA wird nun auf den Textkorpus aus den neuen Dokumenten angewendet. Dies soll verhindern, dass, auch wenn ein Artikel ein bestimmtes Thema behandelt, nicht alle Wörter für diesen temporalen Ausdruck relevant sind.
Im Gegensatz zu Ereignissen, die in der Vergangenheit stattgefunden haben und somit mit einem exakten Datum verknüpft sind, sind Referenzen auf zukünftige Ereignisse häufig nicht exakt. In unterschiedlichen Artikeln können Topics diversen Zeitangaben zugeordnet sein. Anders als bei Referenzen auf vergangene Ereignisse können also Artikel mit ähnlichen Topics, aber unterschiedlichen temporalen Ausdrücken, nicht ohne weiteres verknüpft werden. Aus diesem Grund wurde jeder Ausdruck mit einer Wahrscheinlichkeitsverteilung verknüpft, die in dem gemischten Modell angewendet wird.

1) Ein einzelner Zeitpunkt (z.B. „im Jahr 2020“) wurde mit einer Gaußschen Normalverteilung verknüpft.
2) Ein Enddatum (z.B. „Ende des Jahres 2020“) wurde mit einer wachsenden Exponentialverteilung verknüpft.
3) Ein Anfangsdatum (z.B. „Anfang des Jahres 2020“) wurde mit einer abnehmenden Exponentialverteilung verknüpft.
4) Eine Periode (z.B. „von 2015 - 2025“) wurde mit einer Gleichverteilung verknüpft.

Weiterhin wurden ebenfalls nur der Satz, der den temporalen Ausdruck beinhaltet, sowie der vorherige und darauffolgende Satz betrachtet.

4 Konzept zur Analyse vergangenheitsbezogener Informationen

Yeung und Jatowt verwenden LDA zur Erkennung von Topics im Textkorpus. Anschließend verwenden sie die gewonnenen Daten um weitere Untersuchungen durchzuführen.

4.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation von David M. Blei, Andrew Y. Ng und Michael I. Jordan, im Folgenden mit LDA abgekürzt, ist ein generatives Wahrscheinlichkeitsmodell eines Korpus. Die Korpuselemente, auch Dokumente genannt, können sowohl Bild- als auch Textdokumente sein.
Der Grundgedanke ist, dass die Dokumente aus zufällig gewählten zugrundeliegenden Topics bestehen. Die Topics wiederum sind durch eine Verteilung der sie umfassenden Worte charakterisiert. [2, S. 996]
Im Folgenden wird LDA an Textkorpora beschrieben.

LDA Modell Ein LDA Modell repräsentiert Dokumente als Anhäufung von Topics, die Worte mit einer bestimmten Wahrscheinlichkeit beinhalten. Demnach wird ein Dokument auf folgende Weise erstellt:

- Wähle die Anzahl N von Wörtern, die das Dokument d enthält (z.B. gemäß einer Poisson-Verteilung).
- Wähle eine Menge von Topics aus einer Auswahl von K Topics. Diese kommen in dem Dokument d jeweils mit einer bestimmten Wahrscheinlichkeit vor.
- Generiere jedes Wort w_i in dem Dokument:
  - Wähle ein Topic z gemäß der multinominalen Verteilung, aus dem das Wort stammen soll.
  - Wähle ein spezifisches Wort aus der multinominalen Verteilung der Worte des Topics.

LDA versucht nun mit Hilfe von Backtracking K Topics zu finden, die mit einer hohen Wahrscheinlichkeit die Dokumente des Textkorpus erzeugt haben. [4]

Collapsed Gibbs Sampling Das sogenannte Training ist der Prozess aus einem Textkorpus ein solches Modell zu erzeugen. Für das Training wird ein gewisser prozentualer Anteil des Textkorpus herangezogen. Ist das Training beendet kann, der Rest des Textkorpus mit Hilfe der geschätzten Verteilung untersucht werden. Eine Möglichkeit ein solches Modell zu erzeugen ist das sogenannte Collapsed Gibbs Sampling. [4]

- Ordne jedem Wort w in jedem Dokument d eines der K Topics zu.
- Diese zufällige Zuordnung repräsentiert bereits eine Verteilung aller Topics über die Dokumente und aller Wörter über die Topics. Diese Repräsentation ist natürlich nicht sonderlich gut.
- Um die Zuordnung zu verbessern nehme jedes Dokument d.
  - Nehme jedes Wort w aus d.
    - Berechne das Verhältnis $P(z|d)$ von Wörtern, die diesem Dokument d zugeordnet sind, zu den Wörtern des Dokuments, die dem Topic z zugeordnet sind.
      Berechne das Verhältnis $P(w|z)$ von Dokumenten, denen das Topic z zugeordnet wurde, zu den Dokumenten, in denen das Wort w vorkommt.
      Berechne $P(z|d) \cdot P(w|z)$. Dies entspricht der Wahrscheinlichkeit, dass ein Wort w in einem Dokument d dem Topic z zugeordnet wird. Ordne dem Wort w ein neues Topic z gemäß der berechneten Verteilung zu. Dies beinhaltet die Annahme, dass alle vorhergehenden Berechnungen (bis auf die des aktuellen Wortes) korrekt sind.
- Wiederhole diesen Schritt genügend oft. Man erhält einen annähernd stabilen Zustand mit einer Zuordnung $f : w \to z$.

Angewendet auf den Rest des Textkorpus erhält man:
Ein Topic z wird mit der Wahrscheinlichkeit P_z einem Dokument d zugeordnet, wobei

$$P_z = \frac{|(z,w)| + \eta}{|(z)| + V\eta} \cdot \frac{|(d,z)| + \alpha}{|(d)| + K\alpha}$$

mit $|(z,w)|$ = Anzahl, wie oft das Wort w Topic z zugeordnet ist,
$|(z)|$ = Anzahl, wie viele Wörter Topic z zugewiesen sind,
$|(d,z)|$ = Anzahl, wie viele Wörter in d Topic z zugeordnet sind,
$|(d)|$ = Anzahl, der Wörter in d
und V = die Größe des Vokabulars.

4.2 Einbezug temporaler Ähnlichkeit

Neben den durch Anwendung von LDA gewonnenen Wahrscheinlichkeitsverteilungen können aus den erhobenen Daten $P(y)$, $P(p)$ und $P(p,y)$ für jedes in der Suchanfrage verwendetes Land gewonnen werden. y ist hierbei das referenzierte Jahr eines Dokuments und p das Publikationsjahr des Dokuments. $P(p,y)$ ist demnach die Wahrscheinlichkeit, dass ein Artikel aus dem Jahr p auf das Jahr y referenziert. Yeung und Jatowt benutzen diese leicht zu berechnenden Wahrscheinlichkeiten zur Herleitung weiterer sinnvoller Gleichungen. [1, S. 1235 f.]
Sei nun $P(z|d)$ die Topicverteilung eines Dokuments und D_y die Menge aller Dokumente, die auf das Jahr y verweisen. Die Topicverteilung eines Jahres ist nun

$$P(z|y) = \frac{1}{|D_y|} \sum_{d \in D_y} P(z|d).$$

Es können nun für jedes Land die bedeutendsten Jahre und die dazugehörigen Topics ausgegeben werden. Hierfür werden lediglich die Dokumente betrachtet, die bei der Suchanfrage mit dem Land in Verbindung gebracht wurden.
Es ist ebenfalls interessant zu untersuchen in welchem Jahr ein Topic besonders häufig genannt wird. Diese Wahrscheinlichkeit ist durch

$$P(p|y,z) = \frac{P(p,y,z)}{P(y,z)} = \frac{P(z|p,y)P(p,y)}{P(z|y)P(y)}$$

gegeben.
Weiterhin lässt sich bei gegebenem Publikations- und Referenzjahr die Verteilung der Topics mit

$$P(z|p,y) = \frac{1}{|D_{p,y}|} \sum_{d \in D_{p,y}} P(z|d)$$

angeben.

5 Konzept zur Analyse zukunftsbezogener Informationen

Um ähnliche Ereignisse sowohl nach Topics als auch nach Zeitspannen zu gruppieren verwenden Yeung und Jatowt ein gemischtes Modell, das nachfolgend in Anlehnung an [5, S. 1261 f.] beschrieben wird.

5.1 Grundmodell zum Gruppieren von Dokumenten

In einem einfachen Modell, ähnlich dem LDA, wird zunächst angenommen, dass ein Dokument d auf folgende Weise generiert wird.

$$P(d) = \sum_{z \in Z} P(z) \prod_{w \in W_d} P(w|z)^{N_{w,d}}$$

Wobei $N_{w,d}$ die Anzahl der Erscheinungen eines Worts w in einem Dokument d repräsentiert. Dies stellt die Gruppierung nach Topics dar.

5.2 Einbezug temporaler Ähnlichkeit

Es sollen nun Ereignisse mit einer ähnlichen Wahrscheinlichkeitsverteilung der Zeitspanne gruppiert werden. $G_d(t)$ ist die Wahrscheinlichkeitsfunktion eines Dokuments d, beschrieben in Abschnitt 3.4, und $G_z(t)$ ist die Wahrscheinlichkeitsfunktion eines Topics z. $H(d|z)$ ist nun die Wahrscheinlichkeit, bei gegebenem Topic z, dass d die Wahrscheinlichkeitsverteilung, beschrieben durch $G_d(t)$, hat. $H(d|z)$ wird wie folgt definiert.

$$H(d|z) = \frac{h(d|z)}{\sum_z h(d|z)}$$

mit

$$h(d|z) = \frac{1}{D_{KL}(G_d||G_z) + 1}$$

Hierbei ist $D_{KL}(G_d||G_z)$ die Kullback-Leibler Divergenz, ein statistisches Maß für die Übereinstimmung einer Wahrscheinlichkeitsverteilung p zu einer Modell- oder Kandidatenverteilung q [8, S. 1].

5.3 Gesamtmodell

Zusammen ergibt dies ein gemischtes Modell, das sowohl die Wahrscheinlichkeitsverteilung der Topics als auch die Wahrscheinlichkeitsverteilung der Zeitspannen mit einbezieht.

$$P(d) = \sum_{z \in Z} P(z)(\prod_{w \in W_d} P(w|z)^{N_{w,d}} \times H(d|z)^{\alpha})$$

Der Parameter α steuert den Einfluss der temporalen Ähnlichkeit.
Mit Hilfe des Expectation-Maximization-Algorithmus (EM-Algorithmus) können die Parameter abgeschätzt werden. In zwei Schritten wird bis zu einer gewissen Konvergenz iteriert. [7]

5.4 Logarithmische Zeitachse

Wie in Abschnitt 6.1 gezeigt wird, sind Referenzen in die nahe Zukunft wesentlich genauer als solche, die in die ferne Zukunft verweisen. Aus diesem Grund verwenden Yeung und Jatowt eine logarithmische Einteilung der Zeitachse mit diskreter Einteilung der Werte mit der Einheit Tag als kleinste Einheit.
Somit sind Tage in der nahen Zukunft auf der Zeitachse weiter auseinandergezogen, während Tage in der ferneren Zukunft näher beisammen liegen.

6 Datenanalyse

Nach der Textanalyse konnten Yeung und Jatowt die daraus gewonnenen Ergebnisse untersuchen. Sie zeigten hierbei, dass ihre Vorgehensweise zu interpretierbaren und sinnvollen Schlussfolgerungen führt.

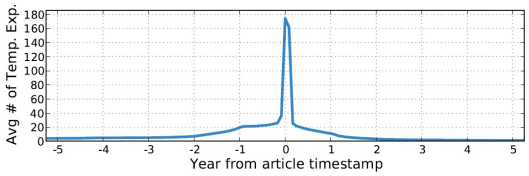
6.1 Allgemeine Feststellungen

Zunächst untersuchten Yeung und Jatowt die gesammelten Dokumente.
Eine ihrer Untersuchungen bezieht sich auf die Verteilung der temporalen Ausdrücke.
Eines der wohl absehbaren Ergebnisse ist, dass mehr Referenzen auf zeitlich nahe Zeitpunkte existieren als auf weiter entfernte. Abbildung 1(a) zeigt, dass die Anzahl temporaler Referenzen mit steigender absoluter Zeitdifferenz zwischen dem Erscheinungsdatum des Artikels und dem Referenzpunkt stetig abnimmt. Es ist auch erkennbar, dass die Kurve sowohl in der Vergangenheit als auch in der Zukunft nach etwa zwei Monaten abrupt fällt. Außerdem ist die Anzahl der temporalen Ausdrücke für Referenzen in die Vergangenheit durchschnittlich höher als die der Referenzen in die Zukunft. [5, S. 1260]
Abbildung 1(b) beschreibt die Verwendung temporaler Ausdrücke bezüglich ihrer Granularität. Ausdrücke wurden in die Granularität von einem Tag, einem Monat und einem Jahr eingestuft. Man kann deutlich erkennen, dass Ausdrücke mit einer feineren Granularität häufiger auf die nahe Vergangenheit und Zukunft referenzieren, grobere Granularitäten häufiger auf zeitlich weiter entfernte Zeitpunkte. [5, S. 1261]

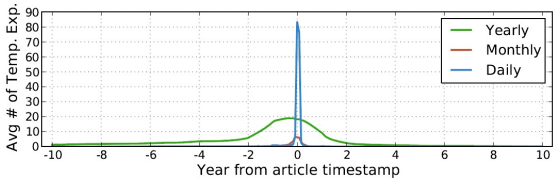
6.2 Analyse von Verweisen in die Vergangenheit

Mit Hilfe der in Abschnitt 4.2 beschriebenen Gleichungen können die gespeicherten Dokumente eingehend auf diverse Merkmale untersucht werden. Die folgenden Abschnitte beschreiben die gewonnenen Erkenntnisse von Yeung und Jatowt.

Verteilung der Referenzen Bereits durch die Verteilung der Referenzen lässt sich zeigen mit welchen Zeitspannen besonders viele Artikel verknüpft sind. In Abbildung 2 ist erkennbar, welche Zeitpunkte in welchem Land besonders erinnert werden. Mit Hilfe unseres Wissens über die Vergangenheit können diesen Zeitpunkten bereits Topics zugeordnet werden.



(a) Verteilung temporaler Referenzen



(b) Verteilung temporaler Referenzen gemäß ihrer Granularität

Abbildung 1. Verteilung temporaler Ausdrücke [5, S. 1261]

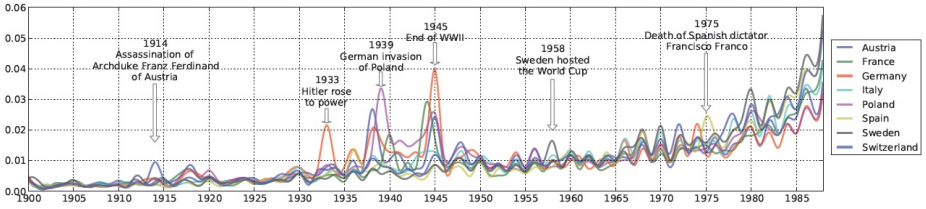


Abbildung 2. Verteilung der in Artikel genannten Jahre einiger europäischer Länder [1, S. 1235]

Signifikante Jahre und Topics Jedes Land ist von eigenen historischen Momenten und Ereignissen geprägt. Tabelle 1 zeigt die drei meistgenannten Jahre und die damit verbundenen Schlagwörter. Nicht nur in den drei in der Tabelle aufgeführten Ländern spielen die Jahre 1939, 1944 und 1945 eine große Rolle. Da sehr viele Nationen im zweiten Weltkrieg involviert waren und dies eine der bedeutendsten Geschehnisse des 20. Jahrhunderts ist, ist dies nicht verwunderlich. Man kann jedoch erkennen, dass innerhalb des zweiten Weltkriegs unterschiedlichen Jahren in unterschiedlichen Ländern eine besondere Bedeutung zugeschrieben wird.

Land	**Jahr**	**Schlagwörter**
Deutschland	1945	war, world, end, day, second, declared, allies, europe, first, empire
	1939	soviet, poland, union, europe, war, eastern, western, czechoslovakia, polish, invaded
	1974	world, cup, final, england, team, first, won, second, win, football
Frankreich	1944	war, world, army, american, battle, soldiers, legion, served, veterans, french
	1940	war, french, german, germany, hitler, occupation, world, resistance, nazi, britain
	1968	killed, people, group, spain, attack, eta, basque, region, year, police
Polen	1939	war, hitler, germany, invasion, britain, invaded, france, german, september, world
	1945	camp, concentration, auschwitz, camps, nazi, death, nazis, sent, january, prisoners
	1980	communist, solidarity, walesa, gdansk, workers, movement, union, leader, government, lech

Tabelle 1. Die drei am häufigsten erwähnten Jahre und die dazugehörigen Schlagwörter einiger europäischer Länder [1, S.1237]

Auslöser Da ermittelt wurde in welchem Land welche Ereignisse besonders häufig referenziert werden ist es auch interessant zu wissen wann und warum dies geschieht. Häufig ist eine Erwähnung mit der Wiederholung eines Ereignisses oder einem Jahrestag verbunden. Für Japan, Deutschland und Polen wurden für Referenzen auf das Jahr 1945 Spitzen in den Jahren 1995 und 2005 gemessen anlässlich des 50. und 60. Jahrestages des Ende des zweiten Weltkriegs. In Japan wurde weiterhin im Jahr 1998 das Jahr 1972 erwähnt. In beiden Jahren war Japan Ausrichter der olympischen Winterspiele. Dieser Zusammenhang kann auch für weitere Ereignisse gezeigt werden.

6.3 Analyse von Verweisen in die Zukunft

Yeung und Jatowt führten einige Fallstudien [5, S. 1262 ff] durch, in denen konkrete Suchanfragen gestellt wurden. Es wurde nach „Germany“, „Toyota“ und „NASA“ gesucht. Im Folgenden wird die Suchanfrage „NASA“ exemplarisch beschrieben.

Fallstudie: NASA Für die Suchanfrage „NASA“ wurden 2499 temporale Ausdrücke, die auf die Zukunft verweisen, extrahiert. Tabelle 2 und Abbildung 3 zeigen die drei häufigsten Cluster für die Suchanfrage.
Das erste Cluster beschreibt die Pläne der NASA 2018 ein bemanntes Raumschiff zum Mond zu senden. Es ist eine eindeutige Spitze im Jahr 2018 zu erkennen.
Das zweite Cluster beschreibt die Bemühungen ein neues Raumschiff zu entwickeln. Es wurde festgestellt, dass dies mehrere verschiedene Projekte betrifft, die in verschiedene Cluster eingeteilt werden sollten. Aufgrund der gleichen Zeitspanne (das Jahr 2014) wurden diese jedoch zusammen gruppiert. Folglich ist es je nach Topic sinnvoll den Einfluss der temporalen Ausdrücke anzupassen.
Das dritte Cluster handelt von einem möglichen Meteoriteneinschlag des Asteroiden Apophis im Jahr 2029. Hier ist die Wahrscheinlichkeitsverteilung weiter gefächert. Dies hat mehrere Gründe. Zunächst hat die Gaußverteilung durch die logarithmische Zeitachse eine höhere Varianz bei weiter entfernten Zeitpunkten. Weiterhin wird in einigen Artikeln der Einschlag früher oder später vermutet. Wie beim zweiten Cluster werden außerdem Artikel mit einbezogen, die mögliche Einschläge weiterer Asteroiden in den Jahren vor und nach 2029 beschreiben.

ID	Anzahl	Häufige Ausdrücke
1	276	moon, space, astronauts, return, mars, agency, president, lunar, programm, new
2	139	space, launch, first, mission, shuttle, agency, flight, spacecraft, orion
3	82	earth, asteroid, space, apophis, amrs, chance, hit, mission, propulsion, scientists

Tabelle 2. Top 3 Cluster für „NASA“

7 Zusammenfassung und Diskussion

Mit Hilfe von relativ einfachen Verfahren ist es möglich umfassende Daten aus einer großen Anzahl von Textdokumenten zu erfassen. Hierfür ist es nicht nötig alle Dokumente zu lesen. Durch Training eines Verfahrens an einem Bruchteil der Dokumente kann der Großteil der Dokumente sehr schnell kategorisiert werden. Die Verfahren lassen sich sowohl zur Informationsgewinnung von Ereignissen in der Vergangenheit als auch von Ereignissen in der Zukunft

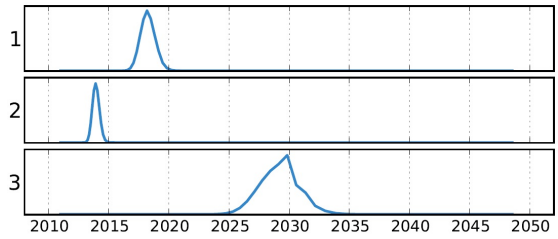


Abbildung 3. Wahrscheinlichkeitsverteilung der drei Cluster für „NASA“ [5, S. 1263]

nutzen. Yeung und Jatowt haben dies anhand ihrer Beispiele gezeigt. Mit Artikeln aus den Jahren 1990 - 2010 untersuchten sie die Vergangenheit von 1900 - 1989 und die nicht allzu ferne Zukunft.
In beiden Fällen ist es jedoch immer noch nicht möglich die gewonnenen Daten ohne jegliches Hintergrundwissen zu interpretieren. Die Verfahren liefern zunächst nur Zeitpunkte / Zeitspannen und die dazugehörigen Schlagworte. Dies liegt daran, dass beliebig viele Topics erkannt aber nicht genannt werden.
Für Ereignisse in der Vergangenheit ist es außerdem möglich die Auslöser der Nennung dieser Ereignisse anzugeben. Dies setzt jedoch auch ein gewisses Maß an Hintergrundwissen voraus.
Die Vorhersage von Ereignissen in der Zukunft ist mit einer Wahrscheinlichkeitsverteilung entlang der Zeitachse verknüpft. Es ist selbstverständlich nicht möglich die Zukunft präzise vorherzusagen. Vielmehr wird ein Bild der Zukunft der heutigen Gesellschaft bzw. der Gesellschaft von 1990 - 2010 gezeigt. Wann genau und ob diese Ereignisse eintreten ist noch unbekannt.

Beide Verfahren liefern gute Ansätze zur Analyse großer Textkörper. Vom sozialwissenschaftlichen Hintergrund dürfte besonders die Analyse der Daten interessant sein, die zu Ergebnissen führt wie die Vergangenheit erinnert und wie sie vergessen wird. Die Daten schildern das Verhalten der kollektiven Erinnerung. Besonders wenn nicht nur englischsprachige Texte mit einbezogen werden kann dies zu interessanten Ergebnissen des kulturellen Verständnisses führen. Dass mit dem Verfahren Wissensplattformen wie Wikipedia oder Suchmaschinen wie Google verdrängt werden ist mehr als unwahrscheinlich. Daten zu einem bestimmten Zeitpunkt zu finden ist inzwischen sehr einfach.
Die Analyse der Erwartungen an die Zukunft bietet jedoch zahlreiche Möglichkeiten, da nicht unbedingt konkrete Ereignisse gesucht sind, sondern vielmehr Trends oder Tendenzen. Diese Informationen können sowohl in der Wirtschaft als auch für den Privatgebrauch genutzt werden, vorausgesetzt man behält in Erinnerung, dass alle Vorhersagen auf einer gewissen Wahrscheinlichkeit beruhen.

Literatur

1. Au Yeung, C.m., Jatowt, A.: Studying how the past is remembered: towards computational history through large scale text mining. In: Proceedings of the 20th ACM international conference on Information and knowledge management. CIKM '11, New York, NY, USA, ACM (2011) 1231–1240
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
3. Cavnar, W.B., Trenkle, J.M.: N-gram-basestext categorization. In: Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. (1994) 161–175
4. Chen, E.: Introduction to latent dirichlet allocation. Webseite (02. Mai 2012) `http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/`.
5. Jatowt, A., Au Yeung, C.m.: Extracting collective expectations about the future from large text collections. In: Proceedings of the 20th ACM international conference on Information and knowledge management. CIKM '11, New York, NY, USA, ACM (2011) 1259–1264
6. Mani, I., Wilson, G.: Robust temporal processing of news. In: In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000). (2000) 69–76
7. Moon, T.K.: The expectation-maximization algorithm. In: IEEE Signal Processing Magazine. (1996) 47–60
8. Shlens, J.: Notes on kullback-leibler divergence and likelihood theory. Webseite (20. August 2007) `http://www.snl.salk.edu/~shlens/`.

Urheberschaftserklärung

Ich versichere hiermit, die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen verwendet zu haben. Alle wörtlichen und sinngemäßen Entlehnungen sind unter genauer Angabe der Quelle kenntlich gemacht. Die Arbeit wurde weder in dieser noch in ähnlicher Form als Prüfungsleistung für eine andere Prüfung eingereicht.

_____	_____
Vorname	Nachname
_____	_____
Ort, Datum	Unterschrift